МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
КАФЕДРА СОЦИОЛОГИИ

С5.я7
К619

В.Ю. КОЛЧИНСКАЯ

АНАЛИЗ ДАННЫХ В СОЦИОЛОГИИ

Учебное пособие

Челябинск
Издательство ЮУрГУ
2006

ББК С5.в6.я7

Колчинская В.Ю. Анализ данных в социологии: Учебное пособие. – Изд-во ЮУрГУ, 2006. – 84 с.

ISBN 5-696-03542-6

Учебное пособие соответствует содержанию курса «Анализ данных в социологии». Содержит материалы для практических занятий и самостоятельной работы студентов с необходимыми пояснениями. Пособие адресовано студентам, обучающимся по специальности «Социология».

Отпечатано с авторского оригинала.

Одобрено учебно-методическим советом исторического факультета ЮУрГУ.

Рецензенты:
Грунт Е.В., *докт. филос. н., профессор УрГУ, г. Екатеринбург,*
Тарасова Ю.Б., *канд. культурологи, доц. ЧГАКиИ, г. Челябинск*

ISBN 5-696-03542-6

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Настоящая книга представляет собой учебное пособие, соответствующее содержанию курса, читаемого автором для студентов исторического факультета Южно-Уральского госуниверситета, обучающимся по специальности «Социология». Целью данного курса является формирование навыков анализа социологической информации. Для реализации этой цели решаются следующие задачи: ознакомление с методологией анализа социологической информации; освоение основных методов и приемов анализа различных видов социологической информации; обучение работе с современными программными системами анализа социологических данных на персональных компьютерах. Дисциплина «Анализ данных в социологии» опирается на знания, полученные студентами при изучении дисциплин: методология и методика социологического исследования, теория измерения в социологии, социальная статистика и другие профессиональные специальные дисциплины.

Акцент в данном пособии делается не столько на теоретическое изложение материала, что прекрасно достигнуто в целом ряде научных публикаций, имеющихся в университетской библиотеке и доступных студентам, сколько не рассмотрении конкретных примеров. Целью такого изложения является желание помочь студентам в решении практических проблем, с которыми они сталкиваются при изучении данного курса и при выполнении собственных исследовательских проектов. Более глубокое же понимание проблем анализа социологических данных может быть достигнуто при изучении литературы, список которой содержится в конце пособия.

Данное пособие состоит из шести частей. В первой рассматриваются некоторые методологические проблемы применения статистических методов для анализа социологических данных, содержание остальных отражает те процедуры, которые необходимо использовать при проведении социологического исследования.

ЧАСТЬ 1. МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ СТАТИСТИЧЕСКИХ МЕТОДОВ В СОЦИОЛОГИЧЕСКОМ ИССЛЕДОВАНИИ

Основные понятия темы

Основные подходы к исследованию социальной реальности: статистический и гуманитарный

Выделяют два основных подхода к исследованию социальной реальности. С точки зрения *статистического*, массовые явления имеют статистический характер, т.е. если изучить достаточно большое количество проявлений изучаемого социального явления, то само явление будет познано. Индивид – представитель некоторой общности, носитель информации о социальном феномене. С точки зрения *гуманитарного* подхода, помощью жестко формализованных методов опроса нельзя заглянуть в глубину социального. Индивид неповторим, и он не проявление социального, а как бы само социальное явление.[1]

Статистическая закономерность

В науке принято выделять две основные формы закономерной связи явлений, отличающиеся по характеру вытекающих из них предсказаний: В законах динамического типа предсказание имеет точный, определенный однозначный вид; в статистических же законах предсказание носит не достоверный, а лишь вероятностный характер.

Статистический подход состоит в мысленном разделении наблюдаемой изменчивости на две части, обусловленные, соответственно, закономерными и случайными причинами, и выявлении закономерной изменчивости на фоне случайной. Вероятностный характер предсказаний в статистических закономерностях обычно бывает обусловлен действием множества случайных факторов, которые имеют место в статистических совокупностях. Статистическая закономерность возникает как результат взаимодействия большого числа элементов, составляющих совокупность, и характеризуют не столько поведение отдельного элемента совокупности, сколько всю совокупность в целом. Проявляющаяся в статистических закономерностях

---

[1] Подробнее об этих подходах можно прочитать в Татарова Г. Г. Методология анализа данных в социологии. – М.: Изд. дом «Стратегия», 1998. – Глава 1. Структура эмпирических данных в социологии; Ядов В. А. Стратегия социологического исследования: Описание, объяснение, понимание социальной реальности. – М.: Академкнига: Добросвет, 2003. – С. 395-397.

"необходимость" возникает вследствие взаимной компенсации и уравновешивания множества случайных факторов[1].

Практические задания

Задание 1. Выбрать некую социальную проблему. В ее границах обозначить аспекты, которые могут быть изучены в рамках: только статистического подхода, только гуманитарного подхода, при сочетании того и другого в разной последовательности.

Задание 2. В рамках аспектов, изучаемых при помощи статистических методов, выделить несколько свойств объектов. Сформулировать эмпирические индикаторы этих свойств. Какими источниками информации можно воспользоваться для изучения данной проблемы?

---

[1] Подробнее об этом см. Толстова Ю.Н. Анализ социологических данных. – М.: Научный мир, 2000. – С. 20-106; Татарова Г. Г. Методология анализа данных в социологии. – М.: Изд. дом «Стратегия», 1998. – Глава 1. Структура эмпирических данных в социологии.

ЧАСТЬ 2. ОРГАНИЗАЦИЯ СТРУКТУРЫ ИСХОДНЫХ СОЦИОЛОГИЧЕСКИХ ДАННЫХ, КЛАССИФИКАЦИЯ ПЕРЕМЕННЫХ

Основные понятия темы

Матрица социологических данных

Первичные данные организуются в виде матрицы, то есть таблицы, в столбцах которой расположены переменные, а в строках – документы. В случае использования метода опроса каждая строка соответствует одной анкете (или одному респонденту).

Переменные

Переменная – это величина, принимающая различные значения в конкретных случаях. В программе она соответствует индикатору, а в анкете – вопросу. Выделяют различные типы переменных.

Ограниченные и неограниченные. Ограниченная может принимать ограниченное число значений. Среди ограниченных выделяют альтернативные и поливариантные переменные. Если альтернативные – допускают возможность выбора только одного варианта, то поливариантные позволяют респонденту выбрать несколько вариантов одновременно.

Примеры[1]

Ограниченная альтернативная переменная:

Отношение к жизни *(выберите один вариант ответа)*

1. никак не могу приспособиться к нынешней жизни;
2. свыкся с тем, что пришлось отказаться от привычного образа жизни; жить, ограничивая себя в большом и малом;
3. мне приходится «вертеться», хвататься за любую возможность заработать, лишь бы обеспечить себе и близким приличную жизнь;
4. мне удалось использовать новые возможности, чтобы добиться большего в жизни;
5. я живу, как и раньше, для меня в последние годы ничего особенно не изменилось;
6. затрудняюсь ответить.

---

[1] В данном разделе примеры переменных взяты из дипломного исследования Бухтиловой М.Н. «Адаптация жителей малого города к социально-экономической ситуации в стране (на примере жителей города Троицка)».

Ограниченная поливариантная переменная:

Наиболее актуальные проблемы *(не более 3 вариантов ответа)*

1. низкий уровень жизни;
2. безработица;
3. рост алкоголизма и наркомании;
4. ухудшение экологической обстановки;
5. низкий уровень культуры в обществе;
6. ухудшение здоровья;
7. плохое медицинское обслуживание;
8. засилье криминала;
9. отмена льгот;
10. нехватка средств для воспитания и обучения детей;
11. высокий уровень цен на товары и услуги;
12. дороговизна жилья;
13. другое____
14. затрудняюсь ответить.

Неограниченная может иметь неограниченное число значений (их не имеет смысла перечислять). Она может быть числовой (в качестве значений имеет числа или строковой (в качестве значений имеет текстовые строки – уникальные наборы символов).

Примеры

Неограниченная строковая переменная

Фамилия, имя, отчество респондента ________________________________

Неограниченная числовая переменная

Возраст (полных лет) ____

Первичные и вторичные. Первичные – замеряемые непосредственно в ходе исследования. Вторичные – рассчитанные с помощью других переменных. Они рассчитываются для решения следующих задач.

Задача *укрупнения интервалов.* Достаточно часто для уменьшения объема информации необходимо сгруппировать значения переменных в интервалы или объединить в более крупные группы.

Задача создания *многомерных таблиц*. В ряде случаев при анализе нам необходимо сформировать группы на многомерной основе, для чего на основе нескольких переменных конструируется комбинационная группировка.

Задача *создания типологий*. При разработке основных показателей исследования, достаточно часто применяется следующая логика: более общие, напрямую не измеряемые признаки разбиваются на простые легко фиксируемые показатели. Однако, затем, на этапе анализа, зачастую необходимо произвести обратную работу: из отдельных простых индикаторов сконструировать общий признак.

Примеры

Укрупнение интервалов

На основе первичной переменной «Доход на одного члены семьи за месяц», принимающей шесть значений, создаем вторичную переменную, принимающую три значения. Значения первичной переменной:

1. До 500 рублей.
2. 500–1200[1] рублей.
3. 1200–2571[2] рублей.
4. 2571–3000 рублей.
5. 3000–4000 рублей.
6. 4000–5500 рублей.

Значения вторичной переменной:

1. Ниже стоимости минимального набора продуктов питания (1200 рублей).
2. Выше стоимости минимального набора продуктов питания, но ниже прожиточного минимума (1200–2571 рублей).
3. Выше прожиточного минимума (2571–5500 рублей).

[1] 1200 рублей на момент проведения исследования составлял размер минимального набора продуктов питания в среднем по Челябинской области.

[2] 2571 рублей на момент проведения исследования составлял прожиточный минимум в среднем по Челябинской области.

Создание многомерных таблиц

С помощью первичных переменных «Пол» и «Возраст» можно сконструировать половозрастные группы. Первичные переменные:

Пол:

1. Мужской.
2. Женский.

Возраст:

1. 18–30 лет.
2. 30–50 лет.
3. Старше 50 лет.

Вторичная переменная

Половозрастные группы:

1. Мужчины 18–30 лет.
2. Мужчины 30–50 лет.
3. Мужчины старше 50 лет.
4. Женщины 18–30 лет.
5. Женщины 30–50 лет.
6. Женщины старше 50 лет.

Создание типологии

На основе двух первичных переменных «Отношение к жизни» и «Оценка материального положения» создается типология «Успешность адаптации». Первичные переменные:

Отношение к жизни:

1. Никак не могу приспособиться к нынешней жизни.
2. Свыкся с тем, что пришлось отказаться от привычного образа жизни; жить, ограничивая себя в большом и малом.
3. Мне приходится «вертеться», хвататься за любую возможность заработать, лишь бы обеспечить себе и близким приличную жизнь.
4. Мне удалось использовать новые возможности, чтобы добиться большего в жизни.
5. Я живу, как и раньше, для меня в последние годы ничего особенно не изменилось.
6. Затрудняюсь ответить.

Оценка материального положения:

1. Полностью обеспечен.
2. Почти полностью обеспечен.

3. Более или менее обеспечен.
4. Мало обеспечен.
5. Бедствую.

На основе различных сочетаний значений первичных переменных создана вторичная переменная «Успешность адаптации», принимающая следующие значения:

1. Полная адаптированность.
2. Незаметная для адаптанта адаптация.
3. Материальный успех с трудностями.
4. Выживание через ограничение потребностей.
5. Отсутствие адаптированности.

Успешность адаптации:

Оценка материального положения:	**Отношение к жизни**					
	Не могу приспособиться	**Отказываюсь от привычек**	**Приходится «вертеться»**	**Новые возможности**	**Живу, как и раньше,**	**Затруднились ответить**
Полностью обеспечен	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация	материальный успех с трудностями	полная адаптированность	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация
Почти полностью обеспечен	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация	материальный успех с трудностями	полная адаптированность	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация
Более или менее обеспечен	незаметная для адаптанта адаптация	выживание через ограничение потребностей	материальный успех с трудностями	полная адаптированность	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация
Мало обеспечен	отсутствие адаптированности	выживание через ограничение потребностей	материальный успех с трудностями	материальный успех с трудностями	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация
Бедствую	отсутствие адаптированности	выживание через ограничение потребностей	материальный успех с трудностями	материальный успех с трудностями	незаметная для адаптанта адаптация	незаметная для адаптанта адаптация

Практические задания

Задание 1. Вам предложен бланк интервью исследования на тему «Ролевая субкультура как ценностно-нормативная система»[1]. Определить виды переменных, которые она содержит.

Задание 2. На основе имеющихся первичных переменных постройте возможные вторичные переменные.

1. С каким настроением Вы просыпаетесь каждое утро?

1. раздражённым; опять будет необходимо решать какие-то проблемы, видеть одних и тех же людей;
2. безразличным; что будет, то будет;
3. радостным; новый день, новые впечатления.

2. Как Вы относитесь к собственным неудачам

1. неудача для меня – это урок на будущее;
2. неудача выбивает меня из колеи.

3. Почему Вы стали играть в ролевые игры? *(Выберите не больше двух ответов.)*

1. Для того, чтобы найти, построить идеальное общество.
2. С целью найти друзей, понимающих тебя людей.
3. Возник интерес к субкультуре, много слышал о ней.
4. Привели друзья.
5. Для разнообразия жизни.
6. Другое (впишите) __________________

4. Для меня ролевики – это *(Выберете один ответ)*

1. Большая семья.
2. Мои друзья.
3. Мои приятели.
4. Компания для времяпровождения.
5. Чужие люди.

---

[1] Этот инструмент был разработан для проведения дипломного исследования Корабельниковой Е.А.

Оцените следующие высказывания, используя 10-и бальную шкалу, где «10» – полностью согласен, «1» не согласен. Вы можете поставить любую оценку от 1 до 10, которая больше всего соответствует Вашему мнению.

5	Если возникнет ситуация, что во время ролевых игр я понадоблюсь моим родителям, я отложу всё и приду им на помощь	
6	Мнение моих родителей о моих действиях, решениях очень важно для меня	
7	Я не обсуждаю с родителями мою личную жизнь	
8	У меня дружная, крепкая семья, которой я горжусь	
9	Я никогда не позволяю себе опаздывать на работу/учёбу	
10	Во время работы/учёбы я стараюсь сконцентрироваться только на выполняемом действии и ни о чем другом не думать	
11	На работу/учёбу я хожу без особого желания, только потому, что так хотят мои родители, жена и т.д.	
12	Рабочий/учебный день пролетает для меня очень быстро, мне редко становится скучно	
13	Моя работа/учёба мне очень нравится, она приносит мне удовольствие	
14	В кругу своих друзей, я всегда являюсь лидером, и не терплю, когда эту роль занимает кто-либо другой	
15	Я считаю, что даже самому близкому другу нельзя доверять свои сокровенные тайны, переживания	
16	Мнение окружающих для меня очень важно, я легко меняю свою точку зрения в сторону большинства	
17	Между мной и моими родителями, присутствует уважение, желание понять, помочь друг другу, даже при конфликтных ситуациях	
18	Мои родители слишком опекают меня, беспокоятся за меня, тем самым сильно ограничивая мои действия, и меня это раздражает	
19	В рабочем/учебном коллективе я чувствую себя как «белая ворона»	

20. В профессиональной/учебной деятельности я сталкиваюсь (сталкивался) с проблемой

1. Отсутствие взаимопонимания с рабочим/учебным коллективом.
2. Отсутствие взаимопонимания с начальником/преподавателем.
3. Разочарование в организации процесса профессиональной/учебной деятельности.

21. Были ли в Вашей жизни ситуации, когда вы уходили (навсегда) с места работы/учёбы, полностью разочаровавшись в ней.

1. Да.
2. Нет.

Как Вы чаще всего проводите свободное время? *Проранжируйте, где «1» – будет чаще всего , «8» – реже всего.*

22	Смотрю телевизор	
23	Общаюсь с родными	
24	Общаюсь с друзьями	
25	Читаю книги, журналы, газеты	
26	Готовлюсь к будущим играм	
27	Слушаю музыку	
28	Хожу на дискотеки	
29	Играю в компьютерные игры	

30. Перечислите, пожалуйста, черты характера, которые свойственны Вам:

____

____

31. Перечислите те качества, которые бы Вы хотели видеть в своём характере:

____

____

32. Какие из перечисленных способов зарабатывания денег для Вас приемлемы?

1. Воровство.
2. Хулиганство.
3. Попрошайничество.
4. Игра на разных видах инструментах на улице.
5. Мелкое мошенничество.
6. Взятки.
7. Азартные игры.
8. Ничего из перечисленного.

Внимательно зачитайте пары суждений и выберете то, которое в большей степени соответствует Вашему мнению. Цифра «3» – означает полное согласие с одним из суждений.

33	Счастливым можно назвать только богатого человека	3	2	1	0	1	2	3	Для счастья богатство не нужно; и в бедности человек может быть счастливым
34	Духовная и физическая близость с любимым человеком – для меня самое главное в жизни	3	2	1	0	1	2	3	Любовь не самое главное для меня
35	Я не представляю свою жизнь без друзей	3	2	1	0	1	2	3	Друзья мне не нужны; я смогу прожить свою жизнь и без них
36	Работа, в которой не предвидится карьерного роста – не для меня	3	2	1	0	1	2	3	Я готов проработать всю жизнь, занимая одну и ту же должность
37	Образование очень важно для меня; это фундамент последующей жизни	3	2	1	0	1	2	3	Образование не играет большой роли в жизни человека
38	На работе я стараюсь сам проявлять инициативу	3	2	1	0	1	2	3	Предпочитаю подчиняться приказам начальства
39	Человеку нужно постоянно развивать новые умения и навыки, т.к. это залог его успешности	3	2	1	0	1	2	3	Для успеха достаточно того, чему научился в молодости
40	Достижение определённого статуса – важнейшая цель моей жизни	3	2	1	0	1	2	3	Я не задумываюсь о достижении определенного положения в обществе
41	Мне бы не хотелось, чтобы кто-нибудь вмешивался в мою жизнь	3	2	1	0	1	2	3	Я считаю допустимым вмешательство в мою личную жизнь других людей
42	Я всегда жертвую своими интересами ради близких мне людей	3	2	1	0	1	2	3	В первую очередь, нужно беспокоиться о себе а потом уже думать о других
43	Каждый человек должен жить и трудиться на благо своей Родины	3	2	1	0	1	2	3	Жить надо, прежде всего для себя и своих близких
44	Для меня очень важно чтобы моя деятельность была одобрена	3	2	1	0	1	2	3	Меня не расстраивает факт непризнания моей деятельности
45	Для меня важно быть уверенным в завтрашнем дне. Я считаю, что жить только настоящим – глупо	3	2	1	0	1	2	3	Нужно жить только настоящим, не задумываясь, что будет завтра

Оцените следующие высказывания, используя 10-и бальную шкалу, где «10» – полностью согласен, «1» полностью не согласен. Вы можете поставить любую оценку от 1 до 10, которая больше всего соответствует Вашему мнению.

Суждение	**46–53. Друзья**	**54–61. Семья**	**61–69. Начальник/преподаватель**	**70–77. Незнакомец**	**78–85. Враг**
Если я в плохом настроении я позволяю себе грубить					
Если меня оскорбили, я всегда отвечу тем же					
При конфликтных ситуациях, я, не задумываясь, применяю физическую силу					
Меня раздражают недостатки других. Я не могу с ними мириться					
Нарушить установленные правила поведения для меня обычное дело					
Я способен на обман, если речь идёт о достижении моих желаний					
Я способен на обман, если от этого будет зависеть моя репутация					
Я способен использовать других людей для достижения личных целей					

86. Мои друзья это *(Выберете один вариант ответа)*

1. Только люди, которые играют в ролевые игры.
2. Мои одноклассники, одногруппники, коллеги по работе и т.д., которые не имеют никакого отношения к ролевым играм.
3. Люди, с которыми мы играем в ролевые игры и люди, которые не имеют к ним никакого отношения.

Каким Вы видите себя через 5 лет? Внимательно зачитайте пары суждений и выберете то, которое в большей степени соответствует Вашему мнению. Для этого Вам нужно выбрать одну из цифр в центральной колонке. Цифра «3» – означает полное согласие с одним из суждений.

87	Продолжаю играть в ролевые игры	3	2	1	0	1	2	3	Появляются совсем другие интересы, играть перестаю
88	Постоянного места работы нет, работаю там, где придётся	3	2	1	0	1	2	3	Есть постоянное место работы
89	Замуж/жениться пока рано, живу один (с родителями, подругой/другом)	3	2	1	0	1	2	3	Выйду замуж/женюсь
90	Живу в своё удовольствие, не задумываясь не о чем	3	2	1	0	1	2	3	Реализую себя в профессиональной деятельности
91	Детей нет	3	2	1	0	1	2	3	Становлюсь отцом/матерью

Расположите (пронумеруйте) следующие сферы жизни в порядке их значимости для Вас. (от 1 до 4, где «1» – самая важная сфера жизни, «4» – наименее важная из перечисленных.)

92	Работа	
93	Досуг	
94	Дружеское общение	
95	Семья	

Каким Вы видите себя через 10 лет? Внимательно зачитайте пары суждений и выберете то, которое в большей степени соответствует Вашему мнению. Для этого Вам нужно выбрать одну из цифр. Цифра «3» – означает полное согласие с суждением.

96	Продолжаю играть в ролевые игры	3	2	1	0	1	2	3	Играть перестаю
97	Постоянного места работы нет	3	2	1	0	1	2	3	Есть постоянное место работы
98	Замуж/жениться, пока рано	3	2	1	0	1	2	3	Выйду замуж/женюсь
99	Живу в своё удовольствие, не задумываясь не о чем	3	2	1	0	1	2	3	Реализую себя в профессиональной деятельности
100	Детей нет	3	2	1	0	1	2	3	Становлюсь отцом/матерью

Внимательно зачитайте пары характеристик и выберете ту, которая в большей степени отражает характер Ваших друзей. Для этого Вам нужно выбрать одну из цифр в центральной колонке. Цифра «3» – означает полное согласие с одной из характеристик.

101	доброта	3	2	1	0	1	2	3	озлобленность
102	оптимизм	3	2	1	0	1	2	3	пессимизм
103	спонтанность	3	2	1	0	1	2	3	рациональность
104	рассеянность	3	2	1	0	1	2	3	организованность
105	веселье	3	2	1	0	1	2	3	серьёзность
106	открытость	3	2	1	0	1	2	3	недоверчивость
107	целеустремленность	3	2	1	0	1	2	3	жизненная неопределённость
108	ум	3	2	1	0	1	2	3	глупость
109	мечтательность	3	2	1	0	1	2	3	реализм
110	медлительность	3	2	1	0	1	2	3	быстрота
111	сила	3	2	1	0	1	2	3	слабость
112	необязательность	3	2	1	0	1	2	3	пунктуальность
113	изменчивость	3	2	1	0	1	2	3	стабильность
114	активность	3	2	1	0	1	2	3	пассивность
115	небрежность	3	2	1	0	1	2	3	аккуратность
116	вспыльчивость	3	2	1	0	1	2	3	спокойность
117	свобода	3	2	1	0	1	2	3	зависимость
118	вежливость	3	2	1	0	1	2	3	грубость
119	доверчивость	3	2	1	0	1	2	3	бдительность
120	трусливость	3	2	1	0	1	2	3	храбрость

121. Перечислите, пожалуйста, какие ещё у Вас есть увлечения помимо ролевой субкультуры ____

Опишите, пожалуйста, человека, с которым бы Вы никогда не хотели заводить знакомств, при виде которого Вы испытываете эмоциональную неприязнь:

122. Как должен выглядеть этот человек ______________________________

____

123. Какие идеи проповедовать ______________________________________

____

Внимательно зачитайте пары характеристик и выберете те, присутствие, которых в человеке делает невозможным Ваше общение с ним. Для этого Вам нужно выбрать одну из цифр в центральной колонке. Цифра «3» – означает полное согласие с одной из характеристик.

124	доброта	3	2	1	0	1	2	3	озлобленность
125	оптимизм	3	2	1	0	1	2	3	пессимизм
126	спонтанность	3	2	1	0	1	2	3	рациональность
127	рассеянность	3	2	1	0	1	2	3	организованность
128	веселье	3	2	1	0	1	2	3	серьёзность
129	открытость	3	2	1	0	1	2	3	недоверчивость
130	целеустремленность	3	2	1	0	1	2	3	жизненная неопределённость
131	ум	3	2	1	0	1	2	3	глупость
132	мечтательность	3	2	1	0	1	2	3	реализм
133	медлительность	3	2	1	0	1	2	3	быстрота
134	сила	3	2	1	0	1	2	3	слабость
135	необязательность	3	2	1	0	1	2	3	пунктуальность
136	изменчивость	3	2	1	0	1	2	3	стабильность
137	активность	3	2	1	0	1	2	3	пассивность
138	небрежность	3	2	1	0	1	2	3	аккуратность
139	вспыльчивость	3	2	1	0	1	2	3	спокойствие
140	свобода	3	2	1	0	1	2	3	зависимость
141	вежливость	3	2	1	0	1	2	3	грубость
142	доверчивость	3	2	1	0	1	2	3	бдительность
143	трусливость	3	2	1	0	1	2	3	храбрость

И в заключение, несколько вопросов о Вас лично:

144. Ваш пол:

1. Мужской. 2. Женский.

145. Сколько Вам полных лет? *(впишите)* ______________

149. Ваше семейное положение:

А) Женат/замужем, (имеется ввиду, только зарегистрированный брак).

Б) Холост/ не замужем.

В) Разведён (а).

150. Скажите, пожалуйста, есть ли у Вас дети:

1. Да. 2. Нет.

146. Ваше основное занятие в настоящее время
1. Рабочий высокой квалификации.
2. ИТР (инженерно-технические работники).
3. Работник торговли, транспорта, сферы услуг.
4. Служащий.
5. Предприниматель.
6. Менеджер.
7. Студент, учащийся.
8. Безработный.
9. Домохозяйка.
10. В отпуске по уходу за ребёнком.
11. Другое(впишите)____

148. Ваш уровень образования
1. Среднее.
2. Среднее специальное (техникум).
3. Незаконченное высшее (не менее 3 курсов).
4. Высшее учебное заведение.
5. Аспирантура.
6. Ученая степень.

151. Укажите, пожалуйста, Ваш основной источник дохода:
1. Помощь родителей.
2. Помощь друзей.
3. Стипендия.
4. Приработки.
5. Постоянная работа.
6. Выигрыши.

152. Как бы Вы охарактеризовали материальное положение своей семьи:
1. Нам не хватает денег даже на еду.
2. Хватает на еду, но покупать одежду мы не можем.
3. Нам хватает на еду и одежду, но мы не можем покупать дорогие вещи.
4. Мы можем покупать дорогие вещи, но не можем покупать всё, что захотим.
5. Мы не ограничены в средствах, полный достаток.

Благодарим за участие в исследовании!

ЧАСТЬ 3. ВЫБОРОЧНЫЙ МЕТОД В СОЦИОЛОГИИ

Основные понятия темы

Генеральная совокупность – объект исследования, очерченный пространственно-временными границами.

Выборочная совокупность – часть объектов генеральной совокупности, выступающих в качестве объектов наблюдения.

Единицы отбора – элементы генеральной совокупности, которые выступают единицами счета в различных процедурах отбора, формирующих выборку.

Единицы наблюдения – элементы сформированной выборочной совокупности, которые непосредственно подвергаются наблюдению. Могут совпадать или не совпадать с единицами отбора.

Репрезентативность – свойство выборки с приемлемой точностью отражать характеристики генеральной совокупности.

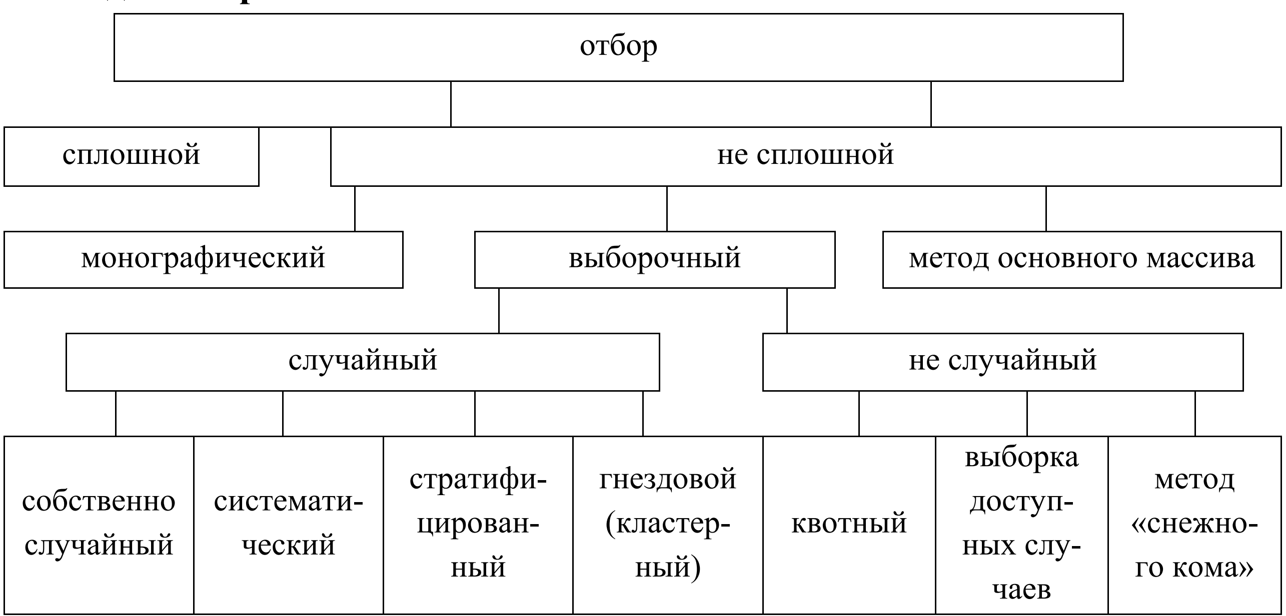
Ошибки наблюдения

Ошибки регистрации – ошибки, отражающие более или менее существенные связи, возникающие в процессе регистрации между объектом, субъектом или условиями наблюдения. *Случайные* ошибки являются результатом погрешностей в результате наблюдения и взаимно погашаются. *Систематические* ошибки приводят к *смещениям.* Причины смещений:

1. Замена намеченных при планировании выборки единиц наблюдения другими, более доступными, которые оказываются неполноценными с точки зрения плана выборки. Это результат некачественной работы интервьюеров.
2. Неполный охват выборочной совокупности. Корректируется с помощью ремонта выборки.
3. Неправильно разработанный план выборки.

Ошибки репрезентативности – расхождение между характеристиками выборочной и генеральной совокупности вследствие не сплошного характера исследования.

Методы отбора



Случайные виды отбора

Случайный отбор – осуществляется, когда каждая единица генеральной совокупности имеет равную вероятность попадания в выборочную совокупность.

Простой случайный отбор может осуществлять как *повторно*, так и *бесповторно*. В первом случае размер генеральной совокупности сохраняется без изменений до конца отбора, что и обеспечивает равную вероятность попадания в выборку всех единиц. Во втором случае вероятность попадания единиц в выборку увеличивается, так как уменьшается размер генеральной совокупности. Если объем генеральной совокупности очень велик по сравнению с объемом выборочной совокупности, этим изменением можно пренебречь. В ходе простого случайного отбора осуществляется выбор единиц наблюдения из перечня объектов генеральной совокупности, называемого *основой выборки*, с помощью таблицы случайных чисел. Требования к основе выборки: полнота – представленность всех единиц генеральной совокупности; отсутствие дублирования; точность – не должна содержать несуществующих единиц; адекватность – соответствие целям исследования; удобство – оформление, облегчающее работу с основой выборки.

Введем следующие обозначения основных показателей:

показатель	генеральная совокупность	выборочная совокупность
объем	N	n
средняя	$\bar{x}$	$\tilde{x}$
доля единиц, обладающих данным значением признака	p	ω
доля единиц, обладающих остальными значениями признака	$q=1-p$	$(1-\omega)$
дисперсия	$\sigma^2_{ген}$	$\sigma^2_{выб}$
дисперсия альтернативного признака	$p(1-p)$	$\omega(1-\omega)$

Ошибка выборки

Средняя ошибка выборки

	Для количественных признаков (ошибка средней)	Для атрибутивных признаков (ошибка доли)
Повторный отбор	$\mu_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$	$\mu_p = \sqrt{\frac{w(1-w)}{n}}$
Бесповторный отбор	$\mu_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$	$\mu_p = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$

Для бесповторного отбора ошибка определяется по формулам, учитывающим величину $\left(1-\frac{n}{N}\right)$. В тех случаях, когда генеральная совокупность очень велика по сравнению с выборочной, эта величина близка к единице, поэтому ею можно пренебречь. Тогда ошибку выборки при бесповторном отборе рассчитывают по формулам для повторного отбора.

Теорема Чебышева–Ляпунова: при достаточно большом количестве наблюдений и при ограниченной дисперсии можно утверждать, что вероятность того, что разница показателей генеральной и выборочной совокупности не превышает заданного предела $t\mu$, стремится к единице. Таким образом, предельная ошибка выборки: $\Delta=t\mu$. Зная как найти среднюю ошибку, получаем:

	Для количественных признаков (ошибка средней)	Для атрибутивных признаков (ошибка доли)
Повторный отбор	$\Delta_{\bar{x}} = t\sqrt{\dfrac{\sigma^2}{n}}$	$\Delta_p = t\sqrt{\dfrac{w(1-w)}{n}}$
Бесповторный отбор	$\Delta_{\bar{x}} = t\sqrt{\dfrac{\sigma^2}{n}\left(1-\dfrac{n}{N}\right)}$	$\Delta_p = t\sqrt{\dfrac{w(1-w)}{n}\left(1-\dfrac{n}{N}\right)}$

Вероятность данного события называют доверительной вероятностью. Величину t называют коэффициентом доверия. Он зависит от уровня доверительной вероятности:

t	1,00	1,96	2,00	2,58	3,00
Φ(t)	0,683	0,950	0,954	0,990	0,997

То есть, с вероятностью 0,954 можно утверждать, что ошибка выборки не превысит удвоенной средней ошибки выборки, с вероятностью 0,997 можно утверждать, что ошибка выборки не превысит утроенной средней ошибки выборки.

С этими величинами тесно связаны следующие характеристики измерения: точность и надежность. **Надежность** отражена в доверительной вероятности: чем больше доверительная вероятность, тем выше надежность измерения. **Точность** оценок отражена в *доверительном интервале* – пределах, в которых с заданной степенью вероятности заключена неизвестная величина оцениваемого параметра. Характеристики выборочной совокупности мы выявляем в результате исследования, характеристики генеральной совокупности можем оценить при помощи доверительного интервала:

$\tilde{x} - \Delta_{\bar{x}} \le \bar{x} \le \tilde{x} + \Delta_{\bar{x}}$
$w - \Delta_p \le p \le w + \Delta p$

то есть, чем больше предельная ошибка, тем выше надежность, но тем ниже точность оценивания характеристик генеральной совокупности. Поэтому зачастую довольствуются несколько меньшей доверительной вероятностью. В социологических исследованиях обычно допустимой считается предельная ошибка доли 0,05.

Определения необходимой численности выборки

Средняя ошибка выборки связана с объемом выборки и степень. Разброса значений признака в генеральной совокупности. Увеличение дисперсии увеличивает ошибку выборки, увеличение объема выборки уменьшает

ошибку выборки. Из формулы предельной ошибки можно рассчитать объем выборки:

	Для количественных признаков (ошибка средней)	Для атрибутивных признаков (ошибка доли)
Повторный отбор	$n = \frac{t^2\sigma^2}{\Delta^2_{\bar{x}}}$	$n = \frac{t^2 w(1-w)}{\Delta^2_p}$
Бесповторный отбор	$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + N\Delta^2_{\bar{x}}}$	$n = \frac{t^2 w(1-w)N}{t^2 w(1-w) + N\Delta^2_{\bar{x}}}$

Доверительная вероятность задается исследователем. Сложность заключается в том, что для расчета объема выборки необходимо знать дисперсию признака, который должен бить измерен в ходе исследования. Эта проблема решается следующими способами:

1. Можно провести пробное обследование, на базе которого определяется величина дисперсии признака, используемая в качестве оценки генеральной дисперсии.

2. Можно использовать данные прошлых обследований, проводившихся в аналогичных целях, то есть дисперсия, полученная по их результатам, используется в качестве оценки генеральной дисперсии.

3. Если нас интересует не среднее значение признака, а доля единиц, обладающих данным значением в совокупности, можно использовать максимально возможную дисперсию, равную 0,25. Тогда мы получим следующие формулы:

Повторный отбор	Бесповторный отбор
$n = \frac{t^2}{4\Delta^2_p}$	$n = \frac{t^2 N}{4N\Delta^2_p + t^2}$

Определяя численность выборки и ее точность, следует учитывать, что чем больше абсолютный объем выборки, тем менее ощутимо влияет на точность результата включение в выборку дополнительных десятков и даже сотен единиц и тем больших затрат требует дальнейшее увеличение точности. Кроме того, объем выборки зависит от предполагаемой группировки объектов, т.е. чем больше будет групп, тем больше должна быть выборка.

Пример

Расчет характеристик случайной выборки

Произведен 10 %-ный случайный бесповторный отбор рабочих для изучения показателей выполнения сменного задания. Выборка дала следующие результаты:

Группы рабочих по проценту выполнения сменного задания	До 100%	100–120%	120–140%	140%–и выше
Численность рабочих	60	250	140	50

Определить

1. средний уровень выполнения сменного задания, гарантируя результат с вероятностью 0,997;
2. долю рабочих, выполняющих задание не менее чем на 120%, гарантируя результат с вероятностью 0,954;
3. необходимый объем выборки при определении доли рабочих, выполняющих сменное задание не менее чем на 120 %, чтобы с вероятностью 0,954 ошибка выборки не превысила 3 %;
4. необходимый объем выборки при определении среднего процента выполнения норм выборки, чтобы с вероятностью 0,950 ошибка выборки не превысила 2 %.

Решение

Задание 1. Генеральная средняя лежит в доверительном интервале: $\tilde{x} - \Delta_{\bar{x}} \le \bar{x} \le \tilde{x} + \Delta_{\bar{x}}$. Для ее определения

1. Рассчитать выборочную среднюю

$$\tilde{x} = \frac{\sum_{i=1}^{k} x_i' f_i}{\sum_{i=1}^{k} f_i} \quad \tilde{x} = \frac{58600}{500} = 117{,}2 \approx 117$$ (см. расчетную таблицу)

2. Определить предельную ошибку выборки $\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$

Для этого нужно:

А) рассчитать дисперсию $\sigma^2 = \frac{\sum_{i=1}^{k}(x_i' - \bar{x})^2 * f_i}{\sum_{i=1}^{k} f_i}$ $\sigma^2 = \frac{134080}{500} = 268{,}16$

(см. расчетную таблицу)

Расчетная таблица

x_i	f_i	x_i'	$x_i'*f_i$	$(x_i' - \bar{x})$	$(x_i' - \bar{x})^2$	$(x_i' - \bar{x})^2 * f_i$
80–100	60	90	5400	–27,2	739,84	44390,4
100–120	250	110	27500	–7,2	51,84	12960
120–140	140	130	18200	12,8	163,84	22937,6
140–160	50	150	7500	32,8	1075,8	53792
Итого Σ	500		58600			134080

Б) определить коэффициент доверия t, используя таблицу значений функции Лапласа, из которой видно, что для *Ф(t)*=0,997 *t*=3,00:

t	1,00	1,96	2,00	2,58	3,00
Ф(t)	0,683	0,950	0,954	0,990	0,997

В) Отношение $\frac{n}{N} = 0.10$, так как в условии сказано, что объем выборочной совокупности составляет 10 % от объема генеральной.

Г) Теперь можем подставить данные в формулу: $\Delta_{\bar{x}} = 3\sqrt{\frac{268}{500}(1-0.1)} = 2.084273 \approx 2{,}08$

3. Зная выборочную среднюю и величину предельной ошибки выборки, можем определить пределы, в которые заключена генеральная средняя:

$\Phi(117 - 2.08 \le \bar{x} \le 117 + 2.08) = 0.997$

$\Phi(114{,}92 \le \bar{x} \le 119{,}08) = 0.997$

Ответ: с вероятностью 0,997 мы можем утверждать, что средний уровень выполнения сменного задания на предприятии не ниже 114,92 % и не выше 119,08 %.

Задание 2. Генеральная доля лежит в доверительном интервале $w - \Delta_p \le p \le w + \Delta p$. Для ее определения

1. Определим долю рабочих, выполняющих задание не менее чем на 120 %, в выборочной совокупности

x_i	f_i	w_i
До 120 %	310	0,62
120 % и более	190	0,38
Итого Σ	500	1,00

Т.е., w_i=0,38

2. Определим предельную ошибку выборки $\Delta_p = t\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$.

А) Для *Ф(t)*=0,954 *t*=2,00 (см. таблицу значений функции Лапласа).

Б) Отношение $\frac{n}{N} = 0.10$, так как в условии сказано, что объем выборочной совокупности составляет 10 % от объема генеральной.

В) Теперь можем подставить данные в формулу: $\Delta_p = 2\sqrt{\frac{0.38(1-0.38)}{500}(1-0.1)} \approx 0.04$

3. Зная выборочную долю и величину предельной ошибки выборки, можем определить пределы, в которые заключена генеральная доля:

$\Phi(0.38 - 0.04 \le p \le 0.38 + 0.04) = 0.954$

$\Phi(0{,}34 \le \bar{x} \le 0{,}42) = 0.954$

Ответ: С вероятностью 0,954 мы можем утверждать, что доля рабочих, выполняющих задание не менее чем на 120 %, на предприятии не ниже 34 % и не выше 42 %.

Задание 3. Объем случайной бесповторной выборки для определения доли единиц, обладающих данным значением признака рассчитывается по формуле $n = \frac{t^2 w(1-w)N}{t^2 w(1-w) + N\Delta^2_{\bar{x}}}$. Нам известна допустимая предельная ошибка выборки (она составляет 3 %). Коэффициент доверия мы можем найти по таблице (t=2.00). Согласно предыдущему исследованию доля рабочих, выполняющих сменное задание не менее, чем на 120 % составляет 0,38. Эти данные мы можем использовать для определения объема выборки. Зная, что 500 человек составляет 10 % от числа рабочих предприятия, мы можем определить объем генеральной совокупности, который равен 5000 человек.

Подставим данные в формулу:

$$n = \frac{2^2 * 0{,}38 * (1-0{,}38) * 5000}{2^2 * 0{,}38 * (1-0{,}38) + 5000 * 0{,}03^2_{\bar{x}}} = \frac{4712}{0{,}9424 + 4{,}5} \approx 866$$

Ответ: необходимый объем выборки составляет 866 человек.

Задание 4. Объем случайной бесповторной выборки для определения среднего значения количественного признака рассчитывается по формуле

$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + N\Delta^2_{\bar{x}}}$. Нам известна допустимая предельная ошибка выборки, она составляет 2 %. Коэффициент доверия мы можем найти по таблице (t=1,96). Согласно предыдущему исследованию дисперсия данного признака составляет 268. Эти данные мы можем использовать для определения объема выборки. Зная, что 500 человек составляет 10 % от числа рабочих предприятия, мы можем определить объем генеральной совокупности, который равен 5000 человек.

Подставим данные в формулу:

$$n = \frac{1{,}96^2 * 268{,}16 * 5000}{1{,}96^2 * 268{,}16 + 5000 * 2^2} = \frac{5150817{,}28}{1030{,}163456 + 20000} \approx 245 .$$

Ответ: необходимый объем выборки составляет 245 человек.

Практические задания

Задание 1. Из общего количества рабочих предприятия была произведена 30 %-ная случайная бесповторная выборка с целью определения затрат времени на проезд к месту работы. Результаты приведены в таблице:

Затраты времени, мин.	Менее 30	30–40	40–50	50–60	60–70
Число рабочих	70	80	200	55	45

Определить:

1. средние затраты времени на проезд к месту работы у рабочих данного предприятия, гарантируя результат с вероятностью 0,997;
2. долю рабочих предприятия, у которых затраты времени на проезд к месту работы составляет 60 мин. и более, гарантируя результат с вероятностью 0,954;
3. необходимый объем выборки для определения среднего времени, затрачиваемого на дорогу, чтобы с вероятностью 0,954 ошибка не превысила 3 мин.;
4. необходимый объем выборки для определения доли рабочих, у которых затраты времени на дорогу составляют 40 мин. и более, чтобы с вероятностью 0,950 ошибка не превысила 5.

Задание 2. На предприятии в порядке случайной бесповторной выборки было опрошено 100 рабочих из 1000 и получены следующие данные об их доходе за октябрь:

Месячный доход, руб.	600–1000	1000–1400	1400–1800	1800–2200
Число рабочих	12	60	20	8

Определить:

1. среднемесячный размер дохода у работников данного предприятия, гарантируя результат с вероятностью 0,997;
2. долю рабочих предприятия, имеющих месячный доход 1400 руб. и выше, гарантируя результат с вероятностью 0,954;
3. необходимую численность выборки, при определении среднего месячного дохода работников предприятия, чтобы с вероятностью 0,954 предельная ошибка выборки не превышала 50 руб.;
4. необходимую численность выборки при определении доли рабочих с размером месячного дохода 1400 руб. и выше, чтобы с вероятностью 0,954 предельная ошибка не превышала 4 %.

ЧАСТЬ 4. АНАЛИЗ ОДНОМЕРНОГО РАСПРЕДЕЛЕНИЯ

Основные понятия темы

Одномерное распределение – это результат группировки единиц совокупности на основе одной переменной. Подобное распределение решает чисто описательные задачи. Его представляют в виде статистических рядов распределения.

Ряды распределения могут быть: а*трибутивными, то есть* построенными по атрибутивному признаку, и *вариационными, то есть* построенными по количественному признаку.

Вариационный ряд будет *дискретным*, если он построен по дискретному признаку, чьи значения обозначены отдельными числами. Если число вариантов дискретного признака слишком велико, а также при анализе вариации непрерывного признака, строятся *интервальные* ряды распределения.

Величины количественного признака и отдельных единиц совокупности различаются между собой. Такое различие в величине признака называется *вариацией*. Ряд распределения характеризуется некоторыми показателями.

Частоты – указывают, сколько раз значение признака встречается в совокупности, их сумма равна количеству исследуемых единиц. В случае, когда группировка осуществляется по альтернативной переменной, исследуемая совокупность делится на непересекающиеся классы, то есть каждый объект может входить только в одну группу. Сумма частот в таком случае равна количеству единиц в совокупности. Если группировка осуществляется по поливариантной переменной, каждый объект может входить в несколько групп. Тогда сумма частот представляет собой количество данных ответов и отличается от количества единиц совокупности.

Помимо частот определяют следующие абсолютные величины:
Число ответивших – сколько человек ответило на данный вопрос.
Число не ответивших – сколько человек не ответило на данный вопрос.
Число опрошенных – сколько всего человек приняло участие в опросе = число ответивших + число не ответивших.
Число ответов – сколько ответов было дано на данный вопрос.

Кроме того, используется такой относительный показатель как п*роценты*, который показывает соотношения пропорций. Используют следующие виды процентов:

% от числа ответивших: единицей анализа в данном случае является человек, ответивший на данный вопрос, то есть не ответившие будут игнорироваться. За 100 % берется число ответивших.

% от числа опрошенных: рассчитывается для того, чтобы определить долю ответивших и не ответивших на данный вопрос. За 100% берется число опрошенных.

% от числа данных ответов: единицей анализа в данном случае выступает не человек, а его ответ. Здесь за 100 % выступает общее число данных ответов.

Поскольку в социологии мы обычно имеем дело с выборочными данными, то перед использованием процентов необходимо учесть статистическую погрешность, то есть ошибку репрезентативности. Таким образом, мы можем говорить о различии между двумя вариантами ответа только в том случае, если разница между ними превышает суммарную ошибку репрезентативности.

Как для абсолютных, так и для относительных частот можно определить *кумулятивные показатели*: накопленная частота (процент) рассчитывается путем суммирования всех частот (процентов), до выбранной категории включительно.

Совокупность в целом характеризуют такие группы показателей как **показатели центра распределения** и показатели вариации. Рассмотрим первую из них.

Мода – это наиболее часто встречающееся значение признака. В интервальном ряду по определению можно установить только модальный интервал. Значение же моды определяется по формуле:

$$\mu o = x_0 + l * \frac{f_{\mu o} - f_{\mu o-1}}{(f_{\mu o} - f_{\mu o-1}) + (f_{\mu o} - f_{\mu o+1})}$$

где

x_0 – нижняя граница модального интервала,

l – величина интервала

$f_{\mu o}$ – частота модального интервала,

$f_{\mu o-1}$ – частота предмодального интервала,

$f_{\mu o+1}$ – частота послемодального интервала

Медиана – это значение признака, обладающее следующим свойством: половина единиц совокупности обладает значением признака большим

либо равным медиане, а вторая половина – меньшим либо равным. Лежит в середине ранжированного ряда.

Для определения медианы нужно сначала рассчитать накопленные частоты по восходящей. Затем – определить порядковый номер медианы по следующей формуле:

$$№\mu e = \frac{n+1}{2},$$ где n – количество единиц в совокупности.

После этого находим накопленную частоту, равную номеру медианы, или первую накопленную частоту, начиная от минимальных значений признака, которая была бы больше номера медианы. Соответствующее ей значение признака и есть медиана. Если номер дробный, то он лежит между двумя единицами совокупности. Тогда медиана равна средней арифметической соседних значений признака.

Для интервального ряда так можно определить только медианный интервал. Медиану рассчитываем по формуле:

$$\mu e = x_0 + l * \frac{\frac{n+1}{2} - s_{\mu e-1}}{f_{\mu e}}$$

где

x_0 – нижняя граница медианного интервала,

l – величина интервала,

n – количество единиц в совокупности,

$s_{\mu e-1}$ – накопленная частота предмедианного интервала,

$f_{\mu e}$ – частота медианного интервала

Средняя арифметическая – это такое значение признака, которое имела бы каждая единица совокупности, если бы общий итог значений был распределен равномерно между всеми единицами совокупности. Эта величина получается от деления суммы всех значений признака на число единиц совокупности. Для сгруппированных данных используется средняя арифметическая *взвешенная*:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$$

где f_i – частота индивидуального значения признака

k – количество градаций признака

Показатели вариации

Размах вариации – разность между максимальным и минимальным значениями признака в изучаемой совокупности

$$R = x_{max} - x_{min}$$

, где x_{max} – максимальное значение признака

x_{min} – минимальное значение признака

Этот показатель прост в расчете, но зависит только от крайних значений признака, поэтому применяется только для однородных совокупностей. Точнее вариацию признака характеризует показатель, основанный на учете колеблемости всех значений признака. Т.к. обобщающей величиной является средняя арифметическая, большинство показателей основано на рассмотрении отклонений от нее индивидуальных значений признака. Таким показателем является с*реднее квадратическое (стандартное) отклонение* (т.к. сумма всех отклонений от средней равна нулю, то возводим их в квадрат). Стандартное отклонение показывает, на сколько в среднем индивидуальные значения признака отличаются от среднего.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 * f_i}{\sum_{i=1}^{k} f_i}}$$

где $\bar{x}$ – среднее значение признака,

x_i – индивидуальное значение признака,

n – общее число единиц наблюдения,

k – количество значений признака,

x_i' – середина интервала

Коэффициент вариации – это отношение стандартного отклонения к средней арифметической, выраженное в процентах: $v = \frac{\sigma}{\bar{x}} * 100\%$. Совокупность считается однородной, если коэффициент вариации не превышает 33% (для распределений, близких к нормальному).

Дисперсия – средняя из квадратов отклонений вариантов значений признака от их средней величины: $\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 * f_i}{\sum_{i=1}^{k} f_i}$.

Помимо дисперсии средней рассчитывается *дисперсия доли.* При наличии двух взаимоисключающих вариантов значений признака говорят о наличии альтернативной изменчивости качественного признака. Эквивалентом такого признака будет переменная, которая принимает значение 1, если обследуемая единица обладает данным признаком, и значение 0, если обследуемая единица не обладает им. К такому виду можно привести любую переменную, выделив группу единиц, обладающих данным значением признака, и группу единиц, обладающих всеми остальными значениями признака. Тогда дисперсия доли будет рассчитана по формуле:

$$\sigma^2 = p(1-p)$$

где p – доля единиц, обладающих данным значением признака

Дисперсия применяется как для оценки рассеяния признака, так и для определения ошибки репрезентативности.

Выбор показателей зависит от исследовательских задач и от уровня, на котором замерен группировочный признак. Для шкал более высокого уровня можно использовать все показатели, которые используются для шкал более низкого уровня, но не все показатели, используемые для шкал более высокого уровня можно использовать для шкал более низкого уровня:

Тип шкалы	Статистические показатели
Номинальная	Количество опрошенных, количество ответивших, количество ответов Частоты, проценты Мода Дисперсия доли
Порядковая	Количество опрошенных, количество ответивших, количество ответов Частоты, проценты, накопленные частоты Мода, медиана Дисперсия доли
Метрическая	Количество опрошенных, количество ответивших, количество ответов Частоты, проценты, накопленные частоты Мода, медиана, средняя арифметическая Размах вариации, стандартное отклонение, коэффициент вариации Дисперсия, дисперсия доли

Примеры

Пример 1. Анализ одномерного распределения по переменной «Проблемы, волнующие респондента».

Переменная ограниченная, поливариантная, номинальная. Данные отсортированы по частоте упоминания, что делает таблицу более наглядной. Из всех видов процентов для анализа используем процент от опрошенных, так как вопрос адресован всей выборочной совокупности. Ошибку выборки можно использовать для анализа значимости различий в количестве выборов между альтернативами. Так, разница между количеством выборов альтернатив «Высокий уровень цен на товары и услуги» и «Низкий уровень жизни» составляет 60%–49%=11%. А суммарная ошибка репрезентативности составляет 6,78%+6,93%=13,71%. Таким образом, мы не можем быть уверены, что первая альтернатива действительно выбирается чаще, чем вторая. Результаты представлены в таблице 4.1.

Таблица 4.1

Проблемы, волнующие респондентов[1]*

Проблемы	Частоты	Проценты			Дисперсия доли	Ошибка выборки
		От опрошенных	От ответивших	От ответов		
Высокий уровень цен	119	60%	60%	20%	0,24	6,78%
Низкий уровень жизни	97	49%	49%	17%	0,25	6,93%
Безработица	54	27%	27%	9%	0,20	6,18%
Нехватка средств для обучения детей	52	26%	26%	9%	0,19	6,11%
Ухудшение здоровья	46	23%	23%	8%	0,18	5,86%
Рост алкоголизма и наркомании	38	19%	19%	7%	0,16	5,47%
Низкий уровень культуры в обществе	36	18%	18%	6%	0,15	5,36%
Дороговизна жилья	34	17%	17%	6%	0,14	5,24%
Отмена льгот	30	15%	15%	5%	0,13	4,98%
Засилье криминала	28	14%	14%	5%	0,12	4,84%
Плохое медицинское обслуживание	24	12%	12%	4%	0,11	4,53%
Ухудшение экологической обстановки	23	12%	12%	4%	0,10	4,45%
Итого	581	291%	295%	100%		

*Объем выборки 200 человек, из них ответило на вопрос 197 человек.

[1] Данные были получены в ходе дипломного исследования Бухтиловой М.Н. «Адаптация жителей малого города к социально-экономической ситуации в стране (на примере жителей города Троицка)».

Мода – высокий уровень цен на товары и услуги.

Пример 2. Анализ одномерного распределения по переменной «Вид вторичной занятости».

Переменная ограниченная, альтернативная, номинальная. Данные отсортированы по частоте упоминания, что делает таблицу более наглядной. Из всех видов процентов для анализа используем процент от ответивших, так как вопрос адресован части выборочной совокупности (тем, кто имеет вторичную занятость). Ошибку выборки можно использовать для анализа значимости различий в количестве выборов между альтернативами. Так, разница между количеством выборов альтернатив «Случайные подработки на основе устной договоренности» и «Формальная регулярная занятость» составляет 50%–23,53%=26,47%, а суммарная ошибка репрезентативности составляет 6,93+5,88=12,81%, что превышает разницу количесва выборов. Следовательно, мы можем утверждать, что в генеральной совокупности на самом деле преобладают случайные подработки на основе устной договоренности. Результаты представлены в таблице 4.2.

Таблица 4.2

Вид вторичной занятости[1]*

Вид вторичной занятости	Частоты	Проценты			Дисперсия доли	Ошибка выборки
		От опрошенных	От ответивших	От ответов		
Случайные подработки на основе устной договоренности	34	17,00%	50,00%	50,00%	0,25	6,93%
Формальная регулярная занятость	16	8,00%	23,53%	23,53%	0,18	5,88%
Регулярная занятость по устной договоренности	15	7,50%	22,06%	22,06%	0,17	5,75%
Формальная нерегулярная занятость	3	1,50%	4,41%	4,41%	0,04	2,85%
Итого:	68	34,00%	100,00%	100,00%		

* Объем выборки 200 человек, из них ответило на вопрос 68 человек.

[1] Данные были получены в ходе дипломного исследования Бухтиловой М.Н. «Адаптация жителей малого города к социально-экономической ситуации в стране (на примере жителей города Троицка)».

Мода – «Случайные подработки на основе устной договоренности».

Пример 3. Анализ одномерного распределения по переменной «Самооценка здоровья».

Переменная ограниченная, альтернативная, порядковая. Данные не отсортированы по частоте упоминания, порядок альтернатив соответствует выраженности признака. Из всех видов процентов для анализа используем процент от опрошенных, так как вопрос адресован всей выборочной совокупности. Так как альтернативы упорядочены в континууме от минимума к максимуму, можно использовать кумулятивные частоты. Последняя кумулятивная частота не равна сумме частот, так как затруднившиеся оценить здоровье не включены в континуум. При расчете порядкового номера медианы также будет использовать количество ответивших на вопрос содержательно, исключая из анализа затруднившихся ответить.

Таблица 4.3

Самооценка здоровья[1]

Оценка	Частоты	Процент от опрошенных	Дисперсия доли	Ошибка выборки	Кумулятивные частоты
Скорее плохое	47	23,50%	0,18	5,88%	47
Очень плохое	13	6,50%	0,06	3,42%	60
Более или менее хорошее	100	50,00%	0,25	6,93%	160
Скорее хорошее	26	13,00%	0,11	4,66%	186
Очень хорошее	12	6,00%	0,06	3,29%	198
Затрудняюсь ответить	2	1,00%	0,01	1,38%	
Итого:	200	100,00%			

Мода – «Здоровье более или менее хорошее».

№ медианы=(198+1)/2=99,5.

Первая кумулятивная частота, больше номера медианы 160, медиана – «Здоровье более или менее хорошее».

[1] Данные были получены в ходе дипломного исследования Бухтиловой М.Н. «Адаптация жителей малого города к социально-экономической ситуации в стране (на примере жителей города Троицка)».

Практические задания

Задание 1. Рассмотрите представленную ниже таблицу одномерного распределения. Рассчитайте статистические показатели, пригодные для данных, замеренных на этом уровне.

Таблица 4.4

Личностные черты, оцениваемые как привлекательные*

Личностные черты	Абсолютные значения
Волевые черты	93
Усидчивость, трудолюбие	58
Доброта, чуткость	36
Коммуникативные навыки	33
Интеллектуальные черты	29
Эмоциональная устойчивость	16
Внешние данные	14
Организованность	12
Нравственные черты	12
Независимость	11
Профессионализм	9
Ответственность, преданность	6
Обаяние, способность оказывать влияние на других	5
Хорошие манеры	5
Оптимизм	4
Чувство юмора	4
Творческий потенциал (оригинальность, инициативность т др.)	4
Чувство собственного достоинства	3
Открытость	2
Практичность (жадность и др.)	2
Жизненный опыт	1
Итого	359

*Количество опрошенных – 466 человек. Не ответили на данный вопрос 195 человека.

Задание 2. Ниже представлены результаты ранжирования ценностей по значимости студентами ЮУрГУ. Выстройте ранжированный ряд ценностей в целом по массиву. Проинтерпретируйте результаты.

Таблица 4.5

Результаты ранжирования ценностей по значимости

(в абсолютных числах)

Ранг	Деньги	Семья	Здоровье	Интерес-ная работа	Дружба	Красота	Любовь	Статус	Самореа-лизация	Рели-гия
1	31	71	101	10	35	2	53	6	21	4
2	58	70	57	17	36	0	53	8	10	4
3	55	52	58	32	35	5	44	6	11	4
4	43	35	36	42	45	2	43	8	21	4
5	35	22	20	62	54	12	30	17	14	4
6	21	20	19	41	46	19	29	17	17	5
7	34	18	20	29	22	13	25	28	24	2
8	16	16	14	39	24	17	20	19	28	6
9	10	13	9	23	20	16	24	28	29	7
10	12	11	10	16	14	16	11	44	17	8
11	11	6	7	16	13	16	3	41	41	6
12	9	6	4	13	6	26	12	36	32	8
13	8	7	2	6	9	43	9	27	40	14
14	11	9	5	11	4	43	4	29	29	14
15	7	7	4	6	3	61	5	25	20	21
16	5	3	4	5	3	48	3	15	11	62
17	4	4	0	2	1	31	2	16	5	197
Итого	370	370	370	370	370	370	370	370	370	370

Алгоритм решения:

1) Рассчитать условный индекс для каждой ценности по формуле средней арифметической взвешенной

2) Проранжировать ценности по величине индекса

3) Рассчитать предельную ошибку выборки для определения значимости различий между ценностями

4) Описать ранжированный ряд в виде текста

5) Сделать социологический вывод относительно полученных результатов

Задание 3. Ниже представлены данные о частоте просмотра телевизора студентами ЮУрГУ, живущими на свои заработки и находящихся на иждивении родственников. Проинтерпретируйте результаты

Таблица 4.6

Частота просмотра телевизора в зависимости от образа жизни

(в абсолютных числах)

Частота просмотра	В целом по массиву	Образ жизни	
		Самостоятельные	Зависимые
Каждый день	148	48	100
4–5 дней в неделю	56	15	41
2–3 дня в неделю	76	29	47
1 день в неделю	19	10	9
Реже одного дня в неделю	42	25	17
Затрудняюсь ответить	29	10	19
Итого:	370	137	233

Алгоритм решения:

1) Рассчитать условный индекс для каждой группы и в целом по массиву по формуле средней арифметической взвешенной

2) Рассчитать предельную ошибку выборки для определения значимости различий между группами

3) Описать результаты в виде текста

4) Сделать социологический вывод относительно полученных результатов

Задание 4. Студентам был предложен вопрос об отношении к телевидению: **Охарактеризуйте ваше отношение к телевидению в целом? (**Вам предложены пары противоположных высказываний. Выберите, к какому из них ближе Ваше мнение и насколько. Отметьте кружком по одной цифре в каждой строке.)

63. Телевидение показывает много захватывающих фильмов	3	2	1	0	1	2	3	На экране слишком много агрессивных фильмов и передач
64. Каждый может выбрать телепередачу по вкусу	3	2	1	0	1	2	3	Телевидение учитывает вкусы не всех зрителей
65. После просмотра телевизора я начинаю чувствовать себя подавленно и угнетенно	3	2	1	0	1	2	3	После просмотра телевизора я нахожусь в приподнятом настроении
66. Телевидение необходимо современному человеку	3	2	1	0	1	2	3	Просмотр телевизора – зря потраченное время
67. Большинство передач и художественных фильмов мне не интересны	3	2	1	0	1	2	3	Мне нравится почти все, что показывают по телевизору
68. Передачи и фильмы показывают реальную жизнь	3	2	1	0	1	2	3	По телевидению показывают искаженную джизнь

В результате обработки данных получили следующие распределения:

Выраженность признака	Пол		Пол		Пол	
	Мужской	Женский	Мужской	Женский	Мужской	Женский
	Телевидение учитывает вкусы всех телезрителей		Телевидение необходимо современному человеку		Просмотр телевизора улучшает мое настроении	
1	40	55	52	45	16	8
2	37	61	32	45	41	31
3	28	20	28	41	27	41
4	26	23	31	35	81	90
5	11	9	17	13	5	11
6	23	14	13	5	7	5
7	14	7	6	5	2	3
Всего:	179	189	179	189	179	189
Выраженность признака	Телевидение позволяет видеть события в динамике		Телевидение показывают жизнь такой, какая она есть		Мне нравится почти все, что показывают по телевизору	
1	20	8	6	8	4	2
2	25	29	13	16	10	10
3	31	36	27	35	20	26
4	39	32	58	68	43	50
5	30	19	33	25	40	29
6	22	26	26	21	33	49
7	12	39	16	16	29	23
Всего:	179	189	179	189	179	189

Необходимо проанализировать полученные результаты.

Алгоритм решения:

1) Рассчитать индексы по формуле средней арифметической взвешенной для каждой из групп

2) Представить результаты в графической форме, располагая позиции в порядке возрастания/убывания различия между ними

3) Рассчитать ошибку выборки, оценить значимость различий между значениями индекса

4) Описать результаты в текстовой форме

5) Сделать социологический вывод относительно полученных результатов

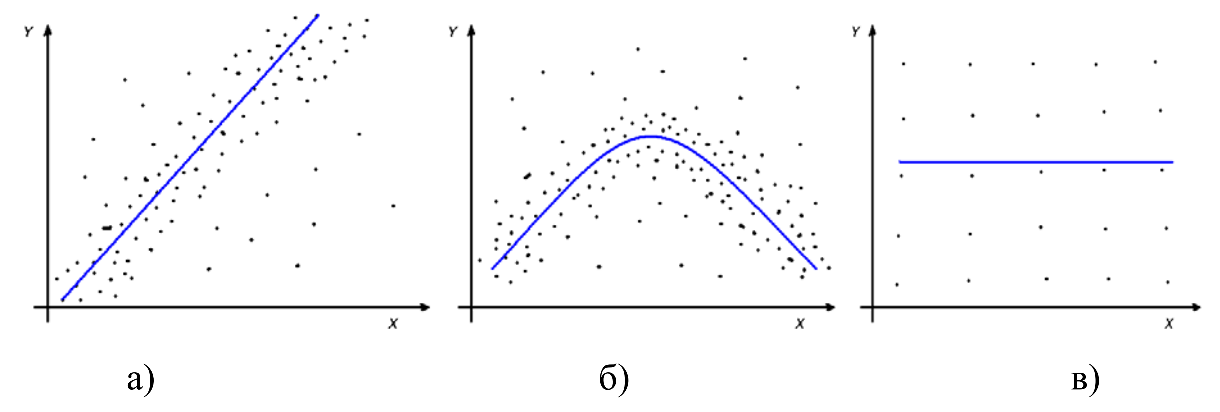
ЧАСТЬ 5. АНАЛИЗ ДВУХМЕРНОГО РАСПРЕДЕЛЕНИЯ

Основные понятия темы

Двухмерное распределение – это распределение единиц совокупности по двум переменным. Его анализ позволяет решать как описательные, так и аналитические задачи. Говоря об описательных задачах, мы имеем в виду, что мы можем охарактеризовать структуру совокупности по двум переменным. Аналитические задачи предполагают установление связи между переменными.

В статистике различают два вида статистической взаимосвязи: функциональную и корреляционную. *Функциональной* называется такая связь между двумя переменными, когда значение *y* однозначно определяется в зависимости от значений *x*. То есть каждому значению *x* соответствует свое значение *y*. Например, функционально связаны общий стаж работы *y* и стаж работы на данном предприятии *x*: $y=ax+b$, где *b* – стаж работы до поступления на предприятие, *a* зависит от особенностей работы. Обычно a=1, но в некоторых случаях (например, при работе на вредном производстве) один год засчитывается за большее количество времени, например, за два, тогда a=2.

В случае *корреляционной* связи значение *y* тоже определяется значением *x*, но не всегда. То есть, каждому значению *x* главным образом соответствует некоторое значение *y*, но не во всех случаях. Графически это можно изобразить следующим образом:



а) б) в)

Рис. 5.1. Пример графиков корреляционной взаимосвязи:
а) линейная прямая; б) нелинейная; в) нет связи

Это происходит в силу того, что на одну и туже переменную влияют несколько факторов. В такой ситуации для того, чтобы увидеть форму взаимосвязи надо рассчитать средний *y* для каждого *x* и рассматривать, как в среднем изменяется *y* при изменении *y*. Таким образом, функциональная

взаимосвязь действует для всех случаев (каждому x или комбинации x_1, $x_2,\ldots,x_n$ соответствует одно значение y). Корреляционная взаимосвязь (каждому x или комбинации x_1, $x_2,\ldots,x_n$ соответствует несколько значений y) выступает только в средних цифрах. Например, если рассматривать зависимость производительности труда от стажа работников, мы увидим, что здесь наблюдается корреляционная зависимость, так как на производительность труда влияют также образование, здоровье, отношение к работе и другие факторы.

Нельзя отождествлять корреляционную и причинно-следственную связь. Наличие корреляции свидетельствует о том, что, либо одно явление является частичной причиной другого, либо оба явления – следствие общих причин. Для выводов о причинно-следственной связи необходимо использовать знание социологической теории.

Выделяют ряд характеристик взаимосвязи. Во-первых, это *сила связи*. Смысл этой характеристики зависит от того, какие коэффициенты корреляции мы используем.

Следующая характеристика – *линейность связи*. Она используется, когда переменные замерены не ниже, чем на порядковом уровне. Связь может быть прямолинейная, если линия, проведенная через средние значения y прямая, и криволинейная, если линия, проведенная через средние значения y кривая. Эти ситуации отращены соответственно на графиках а) и б).

Линейная связь имеет *направление,* то есть ее можно охарактеризовать как прямую или обратную. Прямая связь наблюдается, когда большему значению x соответствует большее значение y. Обратная – когда большему значению x соответствует меньшее значение y.

Помимо всего вышеназванного, при анализе взаимосвязи переменных необходимо оценить *статистическую значимость* связи. Связь считается значимой, если мы можем утверждать, что выявленная на выборочной совокупности закономерность проявляется и в генеральной. Для оценки значимости связи существует целый ряд критериев. Выбор критерия значимости зависит от уровня измерения переменной и коэффициентов взаимосвязи, которые используются.

Анализ двухмерного распределения подчиняется следующей логике.

1. Формулировка гипотезы.

2. Выбор зависимой и независимой переменой.

3. Построение таблицы.

4. Поиск различий по таблице.

5. Оценка статистической значимости различий.

6. Оценка силы и направления связи.

7. Вывод по гипотезе и интерпретация результатов.

Формулировка гипотезы

Мы формулируем гипотезу о взаимосвязи между двумя переменными. Данная гипотеза в статистике обозначается H_1 и называется «*альтернативной гипотезой*». Альтернативная гипотеза о взаимосвязи обычно предполагает и «*нулевую гипотезу*» H_0, о том, что взаимосвязи нет. В результате проверки гипотезы мы должны либо принять нулевую гипотезу и сделать вывод, что связи нет, либо принять альтернативную гипотезу и сделать вывод, что связь есть.

Выбор зависимой и независимой переменой

Формулируя гипотезу, мы предполагаем, что одно переменная – X – будет *независимой*, то есть определяющей. Вторая переменная – Y – будет *зависимой*, то есть определяемой. Напомним, что корреляционная зависимость не говорит о наличии причинно-следственной связи, поэтому названия «зависимая» и «независимая» переменные используются условно.

Построение таблицы

Построение *корреляционной таблицы.* Для анализа корреляционной связи двухмерное распределение представляется в виде таблицы.

Таблица 5.1

Общий вид корреляционной таблицы двух признаков

Y	X						Всего
	x_1	x_2	…	x_j	…	x_k	
y_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1m}	n_1
y_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2m}	n_2
…	…	…	…	…	…	…	…
y_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{im}	n_i
…	…	…	…	…	…	…	…
y_m	f_{k1}	f_{k2}	…	f_{kj}	…	f_m	n_m
Итого	n_1	n_2	…	n_j	…	n_k	n

В этой таблице:

f_{ij} – обозначения внутриклеточных частот, показывают, сколь раз в совокупности встречаются совместно i-е значение Y и j-е значение X.

n_i – маргиналы (итоговые частоты) по *Y,* показывают, сколько раз в совокупности встречается *i*-е значение *Y*.

n_j – маргиналы (итоговые частоты) по *X,* показывают, сколько раз в совокупности встречается *j*-е значение *X*.

N – объем изучаемой совокупности.

Рассмотрев таблицу, мы видим, что каждому значению *X* соответствует не одно определенное значение *Y*, а распределение.

Построение *таблицы средних*. Если зависимая переменная является количественной и замерена с помощью метрической шкалы, для выявления связи между переменными мы можем построить таблицу средних. Для этого независимая переменная, представляется в виде категорий, и для каждой категории рассчитывается среднее по зависимой переменной.

Таблица 5.2

Общий вид таблицы средних

X	Объем групп	$\bar{Y}$	Стандартное отклонение	Коэффициент вариации	Предельная ошибка ±
x_1	n_1	$\bar{y}_1$	σ_1	υ_1	$\pm\Delta_1$
x_2	n_2	$\bar{y}_2$	σ_2	υ_2	$\pm\Delta_2$
…	…	…	…	…	…
x_i	n_i	$\bar{y}_i$	σ_i	υ_i	$\pm\Delta_i$
…	…	…	…	…	…
x_k	n_k	$\bar{y}_k$	σ_k	υ_k	$\pm\Delta_k$
В целом по массиву	n	$\bar{y}_0$	σ_0	υ_0	$\pm\Delta_0$

В этой таблице:

$\bar{y}_i$ – средние значения зависимой переменной *Y* для каждого значения независимой переменной *X*.

υ_i – коэффициент вариации зависимой переменной *Y* для каждого значения независимой переменной *X*.

$\pm\Delta_i$ – предельная ошибка средних значений зависимой переменной *Y* для каждого значения независимой переменной *X*.

n_i – объем групп по независимой переменной *X*.

Поиск различий по таблице

В нашей **корреляционной таблице** приведены абсолютные значения. Для того чтобы сделать предварительные выводы о наличии взаимосвязи признаков, необходимо рассчитать относительные, а именно проценты. В

аналитических целях мы рассчитываем проценты по независимой *(Y)* переменой. Для этого за 100 процентов берем маргинальные частоты по *Y*. Теперь мы можем сравнивать между собой группы, образующие независимую переменную, по такому показателю как доля единиц, обладающих определенным значением зависимой переменной. Если эти доли отличаются, мы можем предположить связь между переменными.

Используя **таблицу средних**, мы можем сравнить между собой группы, образующие независимую переменную, по такому показателю как среднее значение зависимой переменной. Если мы видим различия между этими показателями, то можем предположить связь между двумя переменными.

Оценка статистической значимости различий

Обнаруженные различия между долями и между средними проявляются на обследованной нами совокупности. В случае, когда исследование было сплошным, мы можем этим удовольствоваться. Но чаще всего исследование бывает выборочным, то есть для того, чтобы делать выводы, мы должны проверить, проявляются ли эти различия в генеральной совокупности, которая нас и интересует. Для этого существует ряд критериев.

Для оценки значимости различий между долями или средними используется *t-критерий Стьюдента:*

Сравнение долей

$$t = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}}}$$

где p_1 – доля в первой группе,
p_2 – доля во второй группе,
n_1 – объем первой группы,
n_2 – объем второй группы

Сравнение средних

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

где $\overline{X_1}$ – среднее для первой группы,
σ_1^2 – дисперсия первой группы,
n_1 – объем первой группы (число чел.)
$\overline{X_2}$ – среднее для второй группы,
σ_2^2 – дисперсия второй группы,
n_2 – объем второй группы (число чел.)

Рассчитав значение *t*, мы должны его сравнить с критическими значениями из таблицы критических значений для *t* распределения (см. приложение) и определить вероятность ошибки. Если вероятность ошибки более 0,05, то обычно, такие различия считаются незначимыми, если же вероят-

ность ошибки менее 0,05, то такие различия социологами признаются статистически значимыми.

Для оценки статистической значимости взаимосвязи в целом по таблице мы можем использовать критерий χ^2. С его помощью оценивают значимость отличий данного эмпирического распределения от теоретического распределением, специально рассчитанного таким образом, чтобы в таблице не было взаимосвязи между переменными. Рассчитывает χ^2 по формуле:

$$\chi^2 = \sum_{I=1}^{k}\sum_{j-1}^{m}\left[\frac{(\tilde{n}_{ij}-n_{ij})^2}{\tilde{n}_{ij}}\right]$$

где $\tilde{n}_{ij} = \frac{n_i * n_j}{n}$ – теоретические частоты, то есть те, которые были бы в ячейках таблицы при абсолютной независимости признаков,

n_{ij} – эмпирические частоты,

k, r – число значений признаков

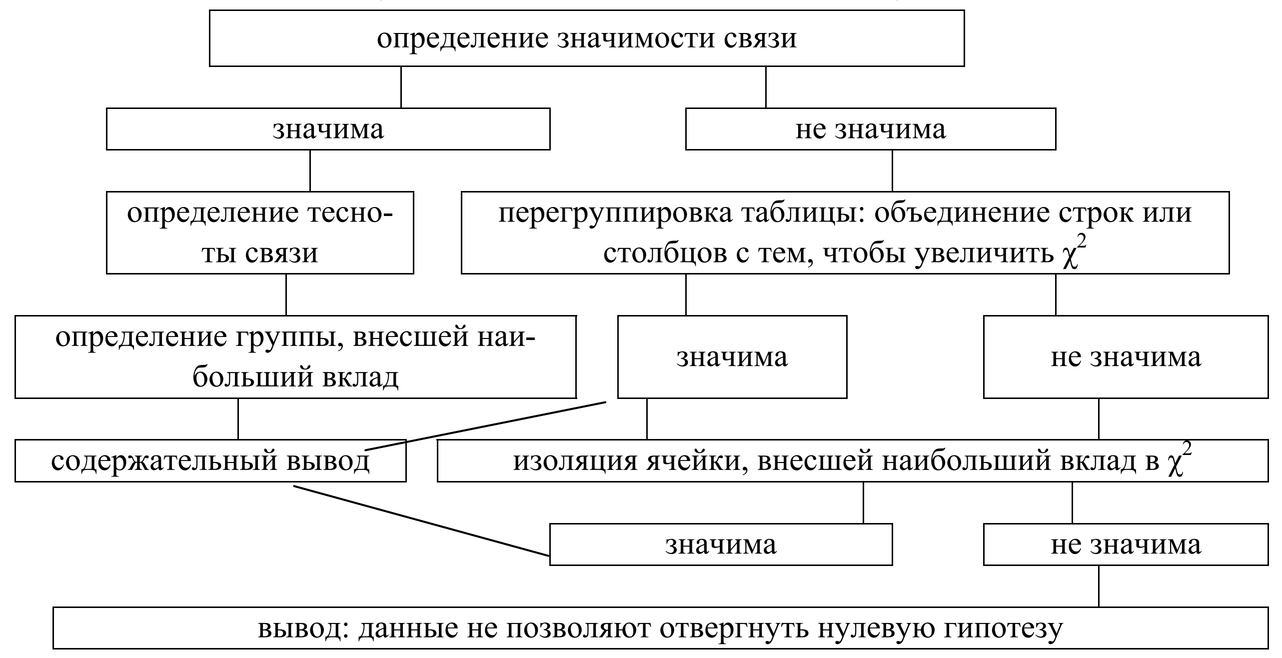
Рассчитав эмпирическое значение χ^2, мы должны сравнить его с табличным критическим значением. Для этого нам потребуется величина df (число степеней свободы):

$$df=(m-1)*(k-1)$$

где *m* – число строк в корреляционной таблице

k – число столбцов в корреляционной таблице

Теперь обращаемся к таблице критических значений. По строчке, соответствующей числу степеней свободы находим критическое значение χ^2 для вероятности ошибки не более 0,05. Если эмпирическое значение превышает критическое, принимаем альтернативную гипотезу. Если не превышает, тогда наши данные не позволяют отвергнуть нулевую гипотезу. Связь не значима. В этом случае можно действовать по следующей схеме:



Оценка силы и направления связи

Оценить силу связи можно при помощи величины коэффициента корреляции по модулю. Обычно силу связи описывают с помощью следующих понятий:

От 0 до 0,3 – связь слабая.

От 0,3 до 0,5 – связь средняя.

От 0,5 до 0,8 – связь сильная.

От 0,8 до 1 – связь очень сильная.

Направление связи оценивается с помощью знака коэффициента корреляции. Положительный коэффициент говорит о наличии прямой связи, а отрицательный – о наличии обратной связи.

Выбор коэффициента корреляции зависит от уровня, на котором замерены данные. Как мы помним, для шкал более высокого уровня можно использовать все показатели, которые используются для шкал более низкого уровня, но не наоборот. Для оценки связи между номинальными переменными, можно использовать только коэффициенты, **основанные на совместном появлении событий.**

Коэффициент Крамера, основан на использовании критерия χ^2:

$$K = \sqrt{\frac{\chi^2}{n * \min(m-1, k-1)}}$$

где n – общее число ответивших на оба вопроса,

m – число строк,

k – число столбцов,

min – надо выбрать наименьшее

Коэффициент Крамера измеряется от 0 до 1, причем, чем ближе коэффициент к 1, тем сильнее связь между двумя переменными. Коэффициент Крамера имеет смысл использовать только в том случае, если связь при помощи χ^2 признана значимой.

Если мы оцениваем связь двух дихотомических признаков, то мы можем использовать некоторые специфические меры связи. Для таких признаков строится корреляционная *таблица размером* 2*2, которая имеет следующий вид:

	B	$\overline{B}$	Σ
A	a	b	a+b
$\overline{A}$	c	d	c+d
Σ	a+c	b+d	n

Для нее рассчитываются следующие коэффициенты:

Коэффициент ассоциации Юла	*Коэффициент контингенции*
$Q = \frac{ad - bc}{ad + bc}$	$\Phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

Эти коэффициенты обладают следующими свойствами:

1. Изменяются в интервале (–1;+1), обращаются в 0 в случае отсутствия связи.
2. Q отражает полную связь (все А=В, но не все В=А), а Ф – абсолютную (все А=В, и все В=А).
3. Q обращается в 1(–1), если хотя бы в одной клетке таблицы частота равна 0, а Ф обращается в 1(–1), если 0 в двух клетках.

Для метрических и порядковых признаков могут использоваться меры, **основанные на принципе ковариации**, то есть на изучении совместных изменений в значениях признаков.

Коэффициент Пирсона можно использовать только для метрических шкал. Он имеет формулу

$$r = \frac{N * \sum XY - \sum X * \sum Y}{\sqrt{\left(N * \sum X^2 - \left(\sum X\right)^2\right) * \left(N * \sum Y^2 - \left(\sum Y\right)^2\right)}}$$

где X – значение независимой переменной,

Y – значение зависимой переменной,

N – объем совокупности

Обе переменные X и Y должны быть в дискретном виде.

Значимость коэффициента можно оценить следующим образом:

Для случая, когда объем совокупности меньше 50, рассчитывается t-критерий по формуле $t = \frac{r}{\sqrt{1 - r^2 * (n-2)}}$

Для случая, когда объем совокупности больше 50, рассчитывается Z-критерий по формуле $Z = \frac{r}{1/\sqrt{n-1}}$

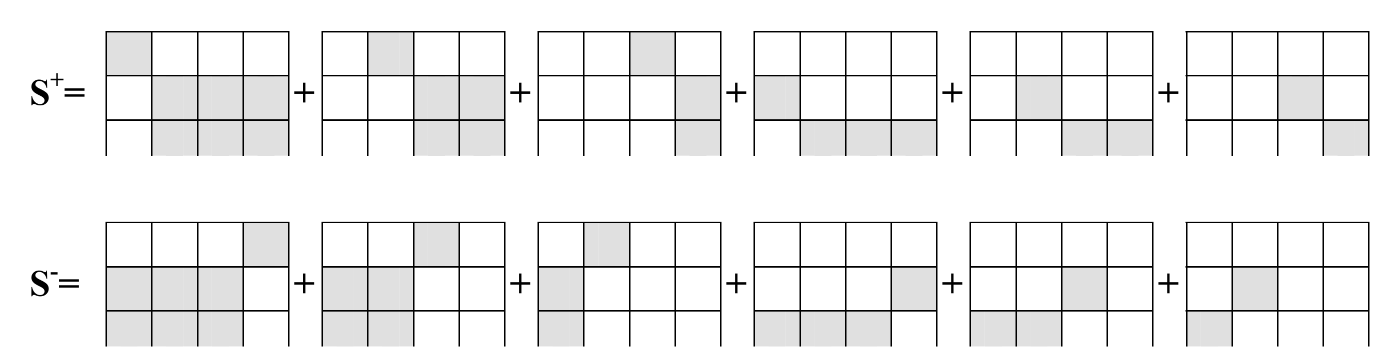
Случаи, когда переменные замерены не ниже, чем на порядковом уровне, удобно оценивать при помощи *γ-коэффициента корреляции Гудмана.* Пусть значения X и Y в корреляционной таблице выписаны в одинаковой последовательности. Тогда γ-коэффициент рассчитывается по формуле:

$\gamma = \frac{S^+ - S^-}{S^+ + S^-}$, где S^+ и S^- – расчетные величины

S^+ рассчитывается по следующей схеме: для каждой ячейки частотной таблицы распределения надо умножить содержимое ячейки на сумму ячеек, находящихся от нее справа и ниже, затем произведения суммируются.

S^- рассчитывается так: для каждой ячейки частотной таблицы распределения надо умножить содержимое ячейки на сумму ячеек, находящихся от нее слева и ниже, затем произведения суммируются.

Наглядно расчет можно представить следующим образом:



γ-коэффициент меняется в интервале от –1 до 1. Положительное значение показывает, насколько вероятно, что при увеличении значения одного признака увеличится значение другого, отрицательный – что при увеличении значения одного признака уменьшится значение другого.

Значимость коэффициента оценивается при помощи *t*-критерия по следующей формуле:

$$t = \gamma * \sqrt{\frac{S^+ + S^-}{n*(1-\gamma^2)}}$$

где *n* – общее число ответивших на оба вопроса

Так же γ-коэффициент имеет смысл сравнивать с коэффициентом Крамера. Если значение коэффициента Крамера значительно больше значения γ-коэффициента, то данная связь носит криволинейный характер, и ее лучше характеризовать при помощи Крамера и процентных различий.

Коэффициент ранговой корреляции Спирмена позволяет определить силу и направление связи между двумя признаками или двумя иерархиями признаков:

$$r_s = 1 - \frac{6*\sum_{i=1}^{l} d_i^2}{l*(l^2-1)}$$

где d_i – разница между парой рангов,
l – количество сравниваемых пар рангов

Для подсчета ранговой корреляции необходимо располагать двумя рядами значений, которые могут быть проранжированы. В случае 1, когда сравниваются два признака, ранжируются индивидуальные значения по

первому признаку, полученные испытуемыми, а затем индивидуальные значения по другому признаку. Если два признака связаны положительно, то испытуемые, имеющие низкие ранги по одному из них, будут иметь низкие ранги и по другому, а испытуемые, имеющие высокие ранги по одному из признаков, будут иметь по другому признаку также высокие ранги. Если же корреляция отсутствует, то все ранги будут перемешаны и между ними не будет никакого соответствия. В случае отрицательной корреляции низким рангам испытуемых по одному признаку будут соответствовать высокие ранги по другому и наоборот.

В случае 2, когда сравниваются две иерархии признаков, ранжируются значения, полученные в двух группах по определенному, одинаковому для двух групп набору признаков. Если эти иерархии связаны положительно, то признаки, имеющие низкие ранги в одной группе, будут иметь низкие ранги и в другой группе и наоборот. При отрицательной корреляции картина обратная: признаки, имеющие высокие ранги в одной группе, имеют низкие ранги в другой.

Значимость коэффициента Спирмена определяется при помощи таблицы критических значений (см. Приложение).

В социологических исследованиях часто объект удается охарактеризовать не по абсолютной, а по относительной интенсивности свойства. Таким образом, известна лишь последовательность, в которой располагаются объекты, то есть каждый объект описывается с помощью рангов по каждому признаку. Еще один *коэффициент ранговой корреляции – коэффициент Кендалла* – строится на основе отношений типа «больше – меньше» и имеет формулу

$$\tau = \frac{2*S}{l*(l-1)}$$

где l – количество сравниваемых пар рангов
S – расчетная величина (см. пример 6)

Значимость коэффициента Кендалла определяется при помощи t-критерия Стьюдента, рассчитываемого по формуле:

$$t = \frac{S}{\sqrt{\frac{1}{18}*l*(l-1)*(2l-5)}}$$

где l – количество сравниваемых пар рангов
S – расчетная величина (см. пример 6)

Интерпретация коэффициента аналогична интерпретации коэффициента Спирмена. Дополнительно следует отметить, что существует несколь-

ко формул расчета коэффициента ранговой корреляции Кендалла, учитывающих число равных рангов.[1]

Когда нам необходимо проанализировать изменение вариации значений одного признака под влиянием другого, имеет смысл использовать *η-коэффициент* (корреляционное отношение). Его использование основано на правиле сложения дисперсий.

Правило сложения дисперсий. Если рассчитать дисперсию признака по всей изучаемой совокупности, то она будет характеризовать вариацию как результат влияния всех факторов, определяющих индивидуальные различия в совокупности. Если же нужно выделить влияние какого-то одного фактора, то совокупность разбивают на группы, положив в основу группировки этот фактор. Выполнение такой группировки позволяет разбить общую дисперсию на две дисперсии, одна из которых будет характеризовать часть вариации, обусловленную влиянием интересующего нас фактора, а вторая – вариацию под воздействием всех прочих факторов.

Вариацию, обусловленную влиянием фактора, положенного в основу группировки, характеризует *межгрупповая дисперсия*, которая является мерой колеблемости групповых средних вокруг общей средней:

$$\delta^2 = \frac{\sum_{j=1}^{m} (\bar{x}_j - \bar{x}_0)^2 * n_j}{\sum_{j=1}^{m} n_j}$$

, где m – число групп

n_j – число единиц в j-ой группе;

$\bar{x}_j$ – средняя арифметическая поj-ой группе;

$\bar{x}_0$ – общая средняя арифметическая

Вариацию, обусловленную влиянием прочих факторов, характеризует в каждой группе *внутригрупповая дисперсия*:

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j}$$

, где x_{ij} – индивидуальное значение признака;

$\bar{x}_j$ – средняя арифметическая поj-ой группе;

n_j – количество единиц в совокупности

По совокупности в целом вариация значений признака под влиянием прочих факторов характеризуется *средней из внутригрупповых дисперсий*:

$$\bar{\sigma}^2 = \frac{\sum_{j=1}^{m} \sigma_j^2 n_j}{\sum_{j=1}^{m} n_j}$$

, где σ_j^2 – внутригрупповая дисперсия;

n_j – количество единиц в совокупности

[1] Подробнее об этом см. Татарова Г. Г. Методология анализа данных в социологии. – М.: Изд. дом «Стратегия», 1998. – Глава 3. Восходящая стратегия анализа данных.

Между общей дисперсией, внутригрупповой дисперсией и средней из внутригрупповых дисперсий существует соотношение, определяемой *правилом сложения дисперсий*: общая дисперсия равна сумме средней из внутригрупповых и межгрупповой дисперсий: $\sigma_0^2 = \overline{\sigma}^2 + \delta^2$

Для оценки влияния фактора нужно рассчитать долю межгрупповой дисперсии в общей – корреляционное отношение:

$$\eta = \sqrt{\frac{\delta^2}{\sigma_0^2}}$$

, η_{min}=0, если δ^2=0, тогда фактор не играет роли;
η_{max}=1, если $\delta^2 = \sigma_0^2$ (σ_l^2=0), тогда вариация полностью обусловлена изучаемым фактором

Проверить значимость η-коэффициента можно при помощи *Z*-критерия следующим образом:

$$Z = \frac{\eta}{1/\sqrt{N-1}}$$

, где *N* – объем совокупности

Вывод по гипотезе и интерпретация результатов

Последний шаг – сделать вывод о гипотезе и провести интерпретацию результатов с точки зрения социологии. Наша гипотеза H_1 (о наличии взаимосвязи) может либо подтвердиться, либо быть опровергнутой (с определенной вероятностью ошибки) и тогда будет иметь смысл принять гипотезу H_0 (об отсутствии связи). Данный вывод мы сделаем на основании проверки статистической значимости различий. Так же (если это подразумевалось гипотезой) надо проверить соответствие гипотетической и эмпирической силы и направленности связи при помощи коэффициента корреляции.

Теперь надо вспомнить, что статистическая и социальная взаимосвязь – разные вещи. Статистическая взаимосвязь определяется по имеющимся данным, а социальная взаимосвязь носит объективный характер. Таким образом, нам надо объяснить вывод о проверке гипотезы с точки зрения социологических знаний, с точки зрения логики причинно-следственных отношений. При этом необходимо учитывать, что статистический вывод может так же объясняться недостаточностью или ненадежностью данных.

Таблица 5.3

Правила оформления различных случаев двухмерного распределения

Независимая	Зависимая		
	Номинальная	Порядковая	Количественная
Номинальная	Таблица: в процентах по независимой переменной *t*-критерий Критерий χ^2 Коэффициент Крамера	Таблица: в процентах по независимой переменной Критерий χ^2 Коэффициент Крамера	Таблица средних значений зависимой переменной η-коэффициент
Порядковая	Таблица: в процентах по независимой переменной Критерий χ^2 Коэффициент Крамера	Таблица: в процентах по независимой переменной γ-коэффициент	Таблица средних значений зависимой переменной. η-коэффициент γ-коэффициент: зависимая количественная в виде интервалов
Количественная	Таблица: количественная в интервалах, процент по независимой переменной. Критерий χ^2 Коэффициент Крамера	Таблица: количественная в интервалах, процент по независимой переменной γ-коэффициент	Таблица средних значений зависимой переменной. η-коэффициент Обе переменные в дискретном виде: коэффициент Пирсона

Таблица 5.4

Коэффициенты и критерии корреляции

Коэффициент или критерий	Познавательные возможности	Измеряется в интервале	Критерий значимости	Формула
Критерии				
χ^2	Определяет разницу между теоретическим и эмпирическим распределением	0...+∞	Таблица критических значений χ^2	$\chi^2 = \sum_{I=1}^{c}\sum_{j-1}^{r}\left[\frac{(\tilde{n}_{ij} - n_{ij})^2}{\tilde{n}_{ij}}\right]$
Стьюдента	Позволяет оценить значимость различий между средними или между долями	–∞...+∞	Таблица интеграла вероятностей	$t = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}}}$ $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Стьюдента	Используется для проверки значимости коэффициентов	–∞...+∞	Таблица интеграла вероятностей	$t = \gamma * \sqrt{\frac{S^+ + S^-}{n * (1-\gamma^2)}}$ $t = \frac{r}{\sqrt{1 - r^2 * (n-2)}}$ $t = \frac{S}{\sqrt{\frac{1}{18} * n * (n-1) * (2n-5)}}$

Z-критерий	Используется для проверки значимости коэффициентов	–∞...+∞	Таблица критических значений Z распределения	$Z = \frac{\eta}{1/\sqrt{N-1}}$ $Z = \frac{r}{1/\sqrt{n-1}}$
Коэффициенты				
Крамера	Позволяет оценить совместную встречаемость значений номинальных переменных	0...1	χ^2	$K = \sqrt{\frac{\chi^2}{N * \min(c-1, r-1)}}$
Спирмена	Сравнение ранжированных рядов	–1...+1	Таблица критических значений Спирмена	$r_s = 1 - \frac{6 * \sum_{i=1}^{n} di^2}{n * (n^2 - 1)}$
Кендалла	Сравнение ранжированных рядов	–1...+1	t-критерий	$\tau = \frac{2 * S}{n * (n-1)}$
γ-коэфициент	Позволяет оценить силу и направленность линейной связи	–1...+1	t-критерий	$\gamma = \frac{S^+ - S^-}{S^+ + S^-}$
η-коэфициент	Позволяет оценить вклад изучаемого фактора в вариацию признака	0..1	Z-критерий	$\eta = \sqrt{\frac{\sigma^2_{межгруппов\ ая}}{\sigma^2_{общая}}}$
Пирсона	Позволяет оценить силу и направление взаимосвязи количественных переменных	–1..+1	Z или t-критерий	$r = \frac{N * \sum XY - \sum X * \sum Y}{\sqrt{\left(N * \sum X^2 - \left(\sum X\right)^2\right) * \left(N * \sum Y^2 - \left(\sum Y\right)^2\right)}}$

Примеры

Пример 1. Оценить взаимосвязь пола и социального положения

Формулировка гипотезы

H_1 – Пол влияет на социальное положение.

H_0 – Пол не влияет на социальное положение.

Выбор зависимой и независимой переменной

Независимая переменная – пол, номинальная, зависимая – социальное положение, номинальная.

Построение таблицы распределения

Таблица 5.5

Социальный статус в зависимости от пола

Статус	Мужчины		Женщины		Всего
	частоты	%	частоты	%	
Рабочие	50	50	10	20	60
Служащие	20	20	10	20	30
Специалисты	30	30	30	60	60
Итого	100	100	50	100	150

Поиск различий по таблице

Среди мужчин рабочие составляют 50 %, специалисты – 30 %. Среди женщин – 60 % специалистов, а рабочих – только 20 %.

Оценка статистической значимости различий

Сравним долю специалистов среди мужчин и среди женщин с помощью *t*-критерия Стьюдента

$$t = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2*(1-p_2)}{n_2}}} = \frac{0,6-0,3}{\sqrt{\frac{0,6*(1-0,6)}{50} + \frac{0,3*(1-0,3)}{100}}} \approx 3,61$$

По таблице критических значений для *t*-распределения Стьюдента находим критическое значение для уровня значимости 0,05, равное 1,96. Эмпирическое значение превышает критическое, значит мужчины и женщины значимо различаются между собой по доле специалистов.

Для оценки статистической значимости взаимосвязи в целом по таблице используем критерий χ^2.

Рассчитываем теоретическое значение для каждой ячейки:

$$\tilde{n}_{ij} = \frac{n_i * n_j}{n}$$

где $\tilde{n}_{ij}$ – теоретическое значение клетки таблицы, находящейся на пересечении строки *i* и столбца *j*

n_i – сумма по строке *i*, n_j – сумма по столбцу *j*,

n – общее число ответивших на оба вопроса

Следующий шаг – расчет разницы между теоретическими и эмпирическими значениями для каждой клетки таблицы по следующей формуле: $\frac{(\tilde{n}_{ij}-n_{ij})^2}{\tilde{n}_{ij}}$. То есть для каждой клетки таблицы нам надо из теоретического значения вычесть эмпирическое (n_{ij} эмпирическое значение можно взять из таблицы частот), разницу возвести в квадрат и разделить на теоретическое значение.

Следующий шаг – подвести сумму по всем клеткам таблицы, получившаяся сумма и даст нам значение $\chi^2 = \sum_{I=1}^{c}\sum_{j-1}^{r}\left[\frac{(\tilde{n}_{ij}-n_{ij})^2}{\tilde{n}_{ij}}\right]$

Последний шаг – сравнение рассчитанного значения χ^2 с табличным критическим значением. Для этого нам потребуется величина df (число степеней свободы) = (число столбцов–1)* (число строк–1). В нашей таблице три строки и два столбца, соответственно df=(3–1)*(2–1)=2. Теперь обращаемся таблице критических значений. По строчке 2 находим ближайшее число к рассчитанному χ^2 и определяем вероятность ошибки. Если вероятность ошибки менее 5% то можно утверждать о статистически значимой взаимосвязи между двумя переменными, иначе следует принять нулевую гипотезу об отсутствии связи.

Построим расчетную таблицу:

n_{ij}	$\tilde{n}_{ij}$	$(n_{ij}-\tilde{n}_{ij})^2$	$\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}$
50	40	100	2.5
20	20	0	0
30	40	100	2.5
10	20	100	5
10	10	0	0
30	20	100	5
			Σ=15

$\chi^2_{эмп}$=15

df=(2–1)(3–1)=2

$\chi^2_{кр}$=5,991

$\chi^2_{эмп} > \chi^2_{кр}$,

т.е., мужчины и женщины значимо отличаются по своему социальному положению

Оценка силы и направления связи

Рассчитаем коэффициент Крамера. В данной таблице количество столбцов меньше, чем количество строк. Поэтому в качестве k и m берем число столбцов – 2.

$$K = \sqrt{\frac{15}{150(2-1)(2-1)}} = 0{,}32$$

Таким образом, мы видим, что связь средняя, ближе к слабой. Направление связи оценить невозможно, так как оба признака замерены на номинальном уровне.

Вывод по гипотезе и интерпретация результатов

Статистический вывод: с вероятностью ошибки меньше 5 % мы можем утверждать, что пол влияет на социальное положение, т.е. данные позволяют нам принять альтернативную гипотезу.

Пример 2. При оценке значимости связи мы не нашли подтверждения альтернативной гипотезе. Необходимо перегруппировать таблицу с целью выявления значимой связи.

Занятость	Возраст				Всего
	До 30 лет	30–50 лет	50–60 лет	60 лет и старше	
Имеющие одно место работы	2,67 10	0.02 13	0,38 14	0,04 23	3,11 60
Имеющие приработок	0,11 6	0,15 22	0,06 26	0.00 36	0,32 90
Безработные	0,20 4	0,14 10	0.11 15	0,05 21	0,50 50
Итого	2,98 20	0,31 45	0,55 55	0,09 80	3,93 200

$\chi^2_{эмп.}$=3,93; $\chi^2_{кр}$=12,592 при df=6, α=0,05

Связь не значима, выбираем, какие строки или столбцы будем объединять. Выделяем столбец, внесший наибольший вклад в χ^2.

Занятость	До 30 лет	Старше 30 лет	Всего
Имеющие одно место работы	2,67 10	0,30 50	2,97 60
Имеющие приработок	0,11 6	0,11 84	0,22 90
Безработные	0,20 4	0,02 46	0,22 50
Итого	3,87 20	0,43 180	200

$\chi^2_{эмп.}$=4,3; $\chi^2_{кр}$=5,991 при df=2, α=0,05

Связь не значима, выбираем ячейку, внесшую наибольший вклад в χ^2 и изолируем ее.

Занятость	До 30 лет	Старше 30 лет	Всего
Имеющие одно место работы	2,67 10	0,30 50	60
Имеющие приработок и безработные	1,14 10	0,12 130	140
Итого	3,81 20	0,42 180	200

$\chi^2_{эмп.}$=4,23; $\chi^2_{кр}$=3,841 при df=1, α=0,05. Связь значима

Пример 3. Оценить взаимосвязь оценки своей жизни и участия в выборах
H_0: Удовлетворенность жизнью не связана с участием в выборах
H_1: Люди, довольные жизнью, чаще участвуют в выборах, чем недовольные
Зависимая переменная – участие в выборах, номинальная, независимая – удовлетворенность жизнью, номинальная. Обе переменные дихотомические.

Таблица 5.6

Участие в выборах в зависимости от удовлетворенности жизнью

	Довольны жизнью	% от ответивших	Не довольны жизнью	% от ответивших	Всего
Голосовали	80	80	40	27	120
Не голосовали	20	20	110	73	130
Итого	100	10	150	10	250

Среди довольных жизнью 80 % участвовали в выборах, а среди недовольных – только 27 %

Рассчитаем эмпирическое значение *t*-критерия Стьюдента:

$$t = \frac{0,80 - 0,27}{\sqrt{\frac{0,80(1-0,80)}{100} + \frac{0,27*(1-0,27)}{150}}} = 181,88,$$ $t_{кр}$=2,576 при α=0,01. Различия значимы.

Полнота связи оценивается с помощью коэффициента Юла:

$$Q = \frac{80*110 - 40*120}{80*110 + 40*20} = 0.83$$

Абсолютность связи оценивается с помощью коэффициента контингенции:

$$\Phi = \frac{80*110 - 40*20}{\sqrt{100/120*150*130}} = 0.52$$

Статистический вывод: с вероятностью ошибки менее 0,01 мы можем утверждать, что люди, довольные жизнью, чаще участвуют в выборах, чем

недовольные. Коэффициент Юла близок к 1, связь полная. Коэффициент контингенции отличается от 1, абсолютная связь не выражена

Пример 4. Проверить гипотезу: ценностное сознание меняется с возрастом.

H_0: Ценностное сознание не зависит от возраста

H_1: Ценностное сознание зависит от возраста

Зависимая переменная – значимые ценности, поливариантная переменная, представляющая собой набор номинальных дихотомических переменных, которые можно упорядочить по частоте выбора, независимая переменная – возраст, порядковая.

Таблица 5.7

Предпочитаемые ценности в зависимости от возраста*

Ценности	Молодежь	Взрослые	Молодежь	Взрослые
	Частоты		Процент от числа ответивших	
Счастливая семейная жизнь	35	123	72,92%	80,92%
Здоровье	35	113	72,92%	74,34%
Материально обеспеченная жизнь	31	142	64,58%	93,42%
Интересная работа	27	56	56,25%	36,84%
Любовь	25	11	52,08%	7,24%
Наличие хороших и верных друзей	21	47	43,75%	30,92%
Карьера	11	6	22,92%	3,95%
Удовольствия	10	16	20,83%	10,53%
Общественное признание	7	52	14,58%	34,21%
Познание нового	7	20	14,58%	13,16%
Активная деятельная жизнь	4	74	8,33%	48,68%
Возможность творчества	2	5	4,17%	3,29%

*Количество ответивших: молодежь – 48, взрослые – 152.

Среди взрослых значимость такой ценности как активная деятельная жизнь отмечают 49 %, а среди молодежи – только 8 %.

Рассчитаем эмпирическое значение *t*-критерия Стьюдента:

$$t = \frac{0,08 - 0,49}{\sqrt{\frac{0,08(1-0,08)}{48} + \frac{0,49*(1-0,49)}{152}}} = -7,09437,$$

$t_{кр}$=2,576 при α=0,01. Различия значимы.

Для сравнения двух ранжированных рядов рассчитаем коэффициент ранговой корреляции Спирмена. Для этого построим расчетную таблицу.

Таблица 5.8

Расчетная таблицы для определения коэффициента Спирмена

Ценности	Молодежь	Взрослые	Молодежь	Взрос лые	d_i	d_i^2
	Частоты		Ранг			
Счастливая семейная жизнь	35	123	1,5	2	–0,5	0,25
Здоровье	35	113	1,5	3	–1,5	2,25
Материально обеспеченная жизнь	31	142	3	1	2	4
Интересная работа	27	56	4	5	–1	1
Любовь	25	11	5	10	–5	25
Наличие хороших и верных друзей	21	47	6	7	–1	1
Карьера	11	6	7	11	–4	16
Удовольствия	10	16	8	9	–1	1
Общественное признание	7	52	9,5	6	3,5	12,25
Познание нового	7	20	9,5	8	1,5	2,25
Активная деятельная жизнь	4	74	11	4	7	49
Возможность творчества	2	5	12	12	0	0
Итого	215	665				114

$r_s = 1 - \frac{6*114}{12*(12^2-1)} = 0,601398601$ $r_{sкр}=0,727$ при α=0,01. $r_{sкр}$ превышает $r_{sэмп.}$, взаимосвязь не значима

Статистический вывод: с вероятностью ошибки менее 0,01 можно утверждать, что предпочитаемые ценности в разных возрастных группах отличаются, но ранжированные ряды не противоположны. Следовательно, имеет смысл сравнивать степень предпочтения отдельных ценностей, применяя, помимо критерия Стьюдента, другие статистические показатели.

Пример 5. Проверить гипотезу: группы респондентов, удовлетворенные содержанием работы, демонстрируют большую удовлетворенность заработной платой, чем неудовлетворенные содержанием работы.

H_0: Удовлетворенность заработной платой и содержанием работы не связаны.

H_1: Удовлетворенность заработной платой и содержанием работы связаны.

Ниже приведены данные об удовлетворенности заработной платой и содержанием работы в группах респондентов с разной степенью адаптиро-

ванности. Сравниваемые переменные замерены на порядковом уровне. Для оценки корреляции удовлетворенности этими сторонами работы рассчитаем коэффициент ранговой корреляции Кендалла.

Таблица 5.9

Расчетная таблица для определения коэффициента Кендалла

Успешность адаптации	Удовлетворенность		Удовлетворенность		S^+	S^-	S
	заработной платой	содержанием работы	заработной платой	содержанием работы			
	Индекс*		Ранг				
Полная адаптированность	0,0	0,4	1	1	4	0	4
Выживание через ограничение потребностей	–0,4	0,0	2	2,5	2	0	2
Незаметная для адаптанта адаптация	–0,5	0,0	3,5	2,5	2	0	2
Материальный успех с трудностями	–0,5	–0,1	3,5	4	1	0	1
Отсутствие адаптированности	–0,8	–0,3	5	5	0	0	0
Всего	—	—	—	—	—	—	9

*Индекс рассчитывается по формуле средней арифметической взвешенной, изменяется в интервале [–1;1].

Для расчета числа согласованных и несогласованных рангов упорядочим ранжированные ряды по первому из признаков. После этого, рассматривая ранги для другого признака, подсчитаем для каждого ранга количество рангов, больше данного, находящихся в таблице ниже него. Это будут значения S^+, то есть число согласованных рангов. После этого, подсчитаем для каждого ранга количество рангов, меньше данного, находящихся в таблице ниже него. Это будут значения S^-, то есть число несогласованных рангов. Затем из каждого количества согласованных рангов вычтем количество несогласованных рангов. Сумма этих величин и будет величина S.

$$\tau = \frac{2*9}{5*(5-1)} = 0,9$$

Для определения статистической значимости рассчитаем t-критерий Стьюдента: $t = \frac{9}{\sqrt{\frac{1}{18}*5*(5-1)*(2*5-5)}} \approx 3,81$ $t_{кр}$=2,576 при α=0,01. Коэффициент значим.

Статистический вывод: с вероятностью ошибки менее 0,01 можем утверждать, что удовлетворенность содержание работы и удовлетворенность заработной платой коррелируют между собой.

Пример 6. Проверить гипотезу: люди, оценивающие свое здоровье как хорошее, легче адаптируются в жизни

H_0: отношение к жизни не зависит от самооценки здоровья

H_1: отношение к жизни зависит от самооценки здоровья

Зависимая переменная отношение к жизни, независимая – самооценка здоровья. Это двухмерное распределение можно представить в виде корреляционной таблицы.

Таблица 5.10

Отношение к жизни в зависимости от самооценки здоровья, чел.

Отношение респондента к нынешней жизни	Оценка здоровья					
	Очень хорошее	скорее хорошее	более или менее хорошее	скорее плохое	очень плохое	Всего
Использовал новые возможности и добился большего	3	7	8	0	0	18
Живу как и раньше, ничего не изменилось	5	6	36	9	2	58
Приходится "вертеться"	2	7	26	10	2	47
Ограничиваю себя во всем	1	3	16	17	4	41
Не могу приспособиться	1	0	8	8	5	22
Всего:	12	23	94	44	13	186

Обе переменные замерены на порядковом уровне, что позволяет определить силу и направленность связи. Для этого рассчитаем γ-коэффициент.

Сначала определим расчетные величины S^+ и S^-

Таблица 5.11

Определение величины S^{+}, чел.

Отношение респондента к нынешней жизни	Оценка здоровья					
	Очень хорошее	Скорее хорошее	Более или менее хорошее	Скорее плохое	Очень плохое	Всего
Использовал новые возможности и добился большего	477*	1001	456	0	—	
Живу как и раньше, ничего не изменилось	530	576	1656	99	—	
Приходится "вертеться"	122	406	884	90	—	
Ограничиваю себя во всем	21	63	208	85	—	
Не могу приспособиться	—	—	—	—	—	
Всего:						6674

477=3(6+7+3+0+36+26+16+8+9+10+17+8+2+2+4+5)

Таблица 5.12

Определение величины S^{-}, чел.

Отношение респондента к нынешней жизни	Оценка здоровья					
	Очень хорошее	Скорее хорошее	Более или менее хорошее	Скорее плохое	Очень плохое	Всего
Использовал новые возможности и добился большего	—	63	200	0	0*	
Живу как и раньше, ничего не изменилось	—	24	504	576	198	
Приходится "вертеться"	—	14	130	290	108	
Ограничиваю себя во всем	—	3	16	153	68	
Не могу приспособиться	—	—	—	—	—	
Всего:						2347

0=0(9+10+17+8+36+26+16+8+6+7+3+0+6+2+1+1)

$$\gamma = \frac{6674 - 2347}{6674 + 2347} = 0{,}49659$$

Для определения значимости коэффициента, рассчитаем t-критерий Стьюдента: $t = 0{,}49659 * \sqrt{\frac{6674 + 2347}{186 * (1 - 0{,}49659^2)}} = 2{,}931082$ $t_{кр}$=2,576 при α=0,01. Коэффициент значим.

Статистический вывод: с вероятностью ошибки менее 0,01 мы можем утверждать, что между переменными существует средняя прямая связь.

Пример 7. В бригаде из восьми человек половина прошли техническое обучение, а половина не прошли. Необходимо определить влияние данного фактора на производительность труда.

H_0: производительность труда не зависит от обучения

H_1: производительность труда зависит от обучения

Независимая переменная – техническое обучение, номинальная, зависимая переменная – производительность труда, количественная.

Имеются следующие данные о производительности труда:

Группа рабочих	Производительность труда			
	№1	№2	№3	№4
прошли техническое обучение	9	8	8	7
не прошли техническое обучение	8	7	7	6

Общие аналитические показатели:

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 * f_i$
6	1	6	1,5	2,25	2,25
7	3	21	0,5	0,25	0,75
8	3	24	0,5	0,25	0,75
9	1	9	1,5	2,25	0,75
Σ	8	60			6

$$\bar{x} = \frac{60}{8} = 7.5$$

$$\sigma^2 = \frac{6}{8} = 0{,}75$$

Внутригрупповая дисперсия прошедших обучение:

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 * f_i$
7	1	7	1	1	1
8	2	16	0	0	0
9	1	9	1	1	1
Σ	4	32			2

$$\bar{x} = \frac{32}{4} = 8$$

$$\sigma^2 = \frac{2}{4} = 0{,}5$$

Внутригрупповая дисперсия не прошедших обучение:

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 * f_i$
6	1	6	1	1	1
7	2	14	0	0	0
8	1	8	1	1	1
Σ	4	28			2

$$\bar{x} = \frac{28}{4} = 7$$

$$\sigma^2 = \frac{2}{4} = 0{,}5$$

Средняя из внутригрупповых дисперсий: $\bar{\sigma}^2 = \frac{0{,}5*4 + 0{,}5*4}{4+4} = 0{,}5$

Межгрупповая дисперсия: $\delta = \frac{(8-7{,}5)^2*4 + (7-7{,}5)^2*4}{8} = 0{,}25$

$$\sigma_0^2 = 0{,}5 + 0{,}25 = 0{,}75$$

$$\eta = \sqrt{\frac{0{,}25}{0{,}75}} \approx 0{,}58\,,$$

Обследование сплошное, поэтому определять значимость не имеет смысла. Таким образом, влияние обучения на производительность труда значительно.

Практические задания

Задание 1. Проверьте исследовательскую гипотезу: люди, имеющие высокий доход не придают значения партийной принадлежности политического лидера

Таблица 5.13

Представление о партийной принадлежности политического лидера в зависимости от уровня дохода респондента, чел.

Партийная принадлежность	Доход			
	Низкий	Средний	Высокий	Всего:
Член влиятельной партии	17	14	8	39
Беспартийный	2	11	13	26
Не имеет значения	19	18	11	48
Затрудняюсь ответить	5	11	2	18
Всего:	43	54	34	131

Задание 2. Проверьте исследовательскую гипотезу: Люди старшего возраста демонстрируют более высокую политическую активность

Таблица 5.14

Формы участия в политической жизни страны в зависимости от возраста, чел.

Формы участия	Возраст				
	18–29	30–44	45–59	60–74	Всего:
Хожу на выборы	40	45	34	35	154
Обсуждаю с друзьями текущую политическую систему в России	17	17	15	13	62
Слежу за политическими процессами в СМИ (чтение газет, журналов, просмотр телепередач)	15	20	9	9	53
Обращаюсь с письмом в СМИ	0	0	1	0	1
Не участвую в политической жизни страны	4	8	8	3	23
Всего:	50	61	47	40	198

Задание 3. Поверьте гипотезу: молодые люди, считающие, что к мнению молодежи в обществе прислушиваются, более активны политически.

Таблица 5.15

Политическая активность молодежи в зависимости от восприятия значимости мнений молодежи (в абсолютных цифрах)

Активность	Восприятие мнения молодежи			Всего:
	Низкая значимость	Средняя значимость	Высокая значимость	
Активные	13	27	17	57
Скорее активные	11	64	21	96
Средние	10	10	4	24
Скорее пассивные	21	26	2	49
Пассивные	29	43	8	80
Всего:	84	170	52	306

Задание 3. Проверьте гипотезу: мужчины и женщины по-разному оценивают крупных политических деятелей прошлого

Таблица 5.16

Популярность политических деятелей прошлого в зависимости от пола
(в абсолютных цифрах)

Политический деятель, вызывающий симпатию	Пол	
	мужской	женский
Александр 1	8	10
Андропов	38	37
Берия	—	—
Бухарин	3	—
Витте	5	8
Иван Грозный	7	2
Екатерина II	12	26
Киров	4	4
Ленин	12	19
Брежнев	2	2
Дзержинский	1	1
Милюков	—	—
Владимир Мономах	1	2
Николай II	7	4
Петр 1	50	69
Свердлов	2	—
Столыпин	21	21
Сталин	25	18
Троцкий	—	3
Хрущев	7	3
Молотов	1	—
Всего:	93	107

Задание 5. Проверить гипотезу: мужчины обладают более высокими притязаниями в сфере заработной платы, чем женщины.

Таблица 5.17

Представление о достойной зарплате в зависимости от пола
(в абсолютных числах)

Достойная зарплата	Пол		Всего:
	Мужской	Женский	
До 5000	2	18	20
5000–7000	15	24	39
7000–10000	12	38	50
10000 и более	55	28	83
Всего:	84	108	192

Задание 6. Проверить гипотезу: люди с высшим образованием обладают более значительными притязаниями в сфере заработной платы, чем люди без высшего образования

Таблица 5.18

Представление о достойной зарплате
в зависимости от уровня образования (в абсолютных числах)

Достойная зарплата	Образование		Всего:
	Среднее специальное	Высшее	
До 5000	12	8	20
5000–7000	22	17	39
7000–10000	16	34	50
10000 и более	22	61	83
Всего:	72	120	192

Задание 7. Проверить гипотезу: учащаяся молодежь старшего возраста демонстрирует большую политическую активность, чем младшие.

Таблица 5.19

Политическая активность в зависимости от возраста
(в абсолютных числах)

Активность	Возраст			Всего:
	До 18 лет	18–20 лет	Старше 20 лет	
Пассивные	16	49	15	80
Скорее пассивные	2	33	14	49
Средние	2	16	6	24
Скорее активные	0	62	34	96
Активные	2	33	22	57
Всего:	22	193	91	306

ЧАСТЬ 6. МНОГОМЕРНЫЙ АНАЛИЗ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

Основные понятия темы

Детерминационный анализ (ДА)[1]

Основная идея ДА – это идея правила, которое можно найти по частотам совпадений или несовпадений событий. Такое правило называется "детерминацией", а математическая теория таких правил – носит название «детерминационный анализ» или ДА.

Люди находят правила (детерминации), наблюдая совпадения либо несовпадения событий. Например, если замечено, что появление А всегда сопровождается появлением В, значит, есть правило "Если А, то В", или, короче, А $\rightarrow$ В. Если А изобразить в виде одного кружка, а В – в виде другого, то кружок А полностью входит в кружок В, как показано на рисунке 1. Это и означает, что имеет место точное правило А $\rightarrow$ В:

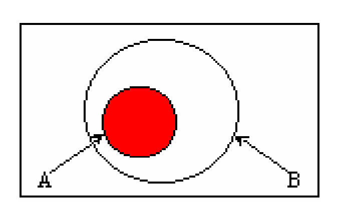


Рис. 6.1. Случай, когда имеется точное правило А $\rightarrow$ В. Кружок А (красный) полностью входит в кружок В. Обрамляющий прямоугольник символизирует весь массив наблюдений.

Идея правила как детерминации тесно связана с идеей предсказания, объяснения. Знание правил позволяет успешно действовать, предвидя результат. В этом причина интереса к правилам. Пример правила, которое может заинтересовать специалиста по предвыборным технологиям: "Если кандидат сделает в таких-то условиях такое-то заявление, его рейтинг в таких-то группах повысится, а в таких-то – станет ниже". Правила – это самая естественная форма знаний, поэтому они нужны всем.

Любое правило имеет две фундаментальные характеристики – *точность и полноту*. Точность правила А $\rightarrow$ В это, по определению, доля случаев В среди случаев А. На рисунке 1 эта доля равна 1 (100%), что и означает, что правило А $\rightarrow$ В предельно точное. Помимо точности есть еще

---

[1] Подробно см. Чесноков С.В. Детерминационный анализ социально-экономических данных. – М.: Наука, 1982.

одна фундаментальная характеристика – полнота. Из рисунка 1 видно, что с помощью правила A → B можно предсказать лишь примерно одну четверть всех случаев появления B. Чтобы применить правило A → B, нужно сначала обнаружить A, и только после этого можно предсказать наличие B. A площадь кружка A составляет примерно одну четверть от площади кружка B. Правило A → B точное, но не полное, его полнота равна примерно одной четверти (25%).

В общем случае полнота правила A → B есть, по определению, доля случаев A среди случаев B. Полнота правила A → B равна точности обратного правила B → A, а точность правила A → B равна полноте обратного правила. При перемене направления стрелки в любом правиле точность и полнота меняются местами.

Неточное правило можно сделать точным. Точных правил не так много. Большинство правил – неточные. Если правило A → B неточное, кружок A не полностью входит в кружок B, как показано на рисунке 2.

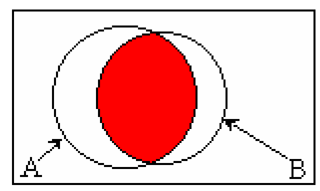


Рис. 6.2. Случай, когда имеется неточное правило A → B. Только часть кружка A (окрашена красным) входит в кружок B.

Если в неточное правило A → B добавить некоторый фактор C, может случиться, что правило AC → B, которое получится в результате, будет точным. Пример такой ситуации показан на рисунке 3.

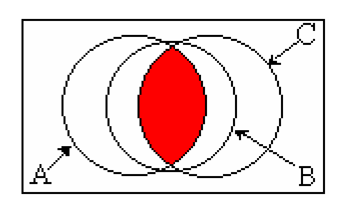


Рис. 6.3. В неточное правило A → B добавлен фактор C. В результате получилось точное правило AC → B. Все случаи, когда имеется сочетание A и C (окрашены красным) оказались внутри кружка B.

Конечно, может случиться, что точность правила AC → B будет еще менее точным, чем первоначальное правило A → B. На рисунке 4 правило AC → B имеет точность, равную нулю.

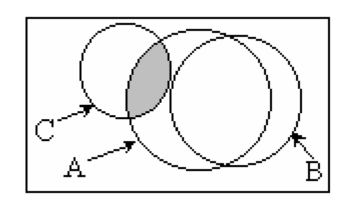


Рис. 6.4. В неточное правило A → B добавлен фактор C. В результате получилось правило AC → B, которое имеет точность, равную нулю. Все случаи, когда имеется сочетание A и C (окрашены серым) оказались вне кружка B.

Чесноков для обозначения того объекта, который является носителем локальной связи, вводится понятие детерминации, обозначаемой a→b. Детерминация определяется как носитель локальной связи или как нечто, задаваемое двумя величинами:

точностью I (a→b) = P(b/a) и

полнотой C (a→b) = P (a /b) (справа стоят относительные частоты).

Факторный анализ[1]

Одна из важных задач статистики – сделать эмпирическую информацию компактной, удобной для анализа. Одним из направлений конденсации информации является факторный анализ признаков.

Основная идея. Индивиды обладают самыми разнообразными признаками, которые не являются независимыми. Связи между ними изучаются с помощью методов корреляционного анализа. Можно предположить, что некоторые признаки образуют группы, каждая из которых отражает определенный аспект сложного явления. При анализе системы признаков мы сталкиваемся с классификацией признаков, т. е. с выявлением групп признаков, имеющих сходный характер изменения при переходе от одного объекта к другому. В частности, ставится задача найти максимально взаимосвязанные группы признаков. Выделяемые группы – это новые, комплексные переменные, называемые *факторами*.

Обоснованная замена большого числа признаков, описывающих объекты наблюдения, меньшим числом комплексных характеристик (факторов) составляет сущность факторного анализа.

Подчеркнем, что факторы не сводятся к некоторым, пусть главным, основным признакам исходного набора. Каждый фактор – это группа взаимосвязанных признаков из упомянутого набора, и вся совокупность входя-

[1] Подробно см. Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. Киев: Наукова думка, 1982; Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989.

щих в него признаков определяет содержательную интерпретацию этого фактора.

Факторный анализ позволяет не только выделить группы наиболее взаимосвязанных признаков, но и отделить несущественные признаки от существенных, оценить их информативность.

В ходе факторного анализа выделяется латентная переменная-фактор, с которой коррелируют первичные переменные. Эти корреляции называются *факторными нагрузками*. Кроме того, рассматривают корреляцию факторов между собой.

Кластерный анализ[1]

Еще одним направлением конденсации информации является классификация объектов. В качестве синонимов для обозначения этой группы методов используют такие термины как «кластерный анализ», «таксономия», «автоклассификация» или (более широко) говорят об использовании методов «распознавания образов». Пусть, матрица данных включает характеристики N объектов по двум количественным признакам (например, стаж работы и зарплата). Откладывая признаки по осям координат, мы можем изобразить все объекты на плоскости в виде N точек: абсцисса – значение стажа, ордината – значение зарплаты данного объекта. В этом случае говорят, что N объектов расположены в двухмерном признаковом пространстве; (по сути, это один из способов изображения двухмерного распределения признаков). Как видно из рисунка, все объекты можно разбить на три группы таким образом, что объекты внутри групп близки между собой (это означает, что они имеют близкие характеристики и по X и по Y), а объекты из разных групп – далеки.

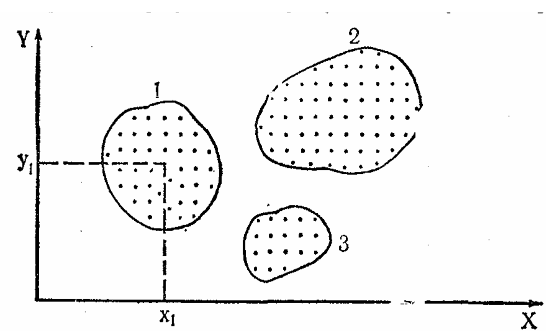


Рис. 6.5. Изображение объектов в пространстве двух признаков

[1] Подробно см. Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. Киев: Наукова думка, 1982; Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989.

Множество близких между собой точек называется *кластером* и при интерпретации результатов рассматривается как некоторый социальный тип. Если имеется k признаков, то говорят, что объекты расположены в k-мерном признаковом пространстве. Если признаков более чем два, то точки уже невозможно изобразить на плоскости. В этом случае группировку можно осуществить с помощью формальных методов.

Результатом работы алгоритмов обычно является разбиение множества объектов на группы в пространстве признаков, заданных исследователем, а также расчет некоторых обобщенных характеристик каждого из кластеров (центр кластера, средние, меры вариации). Существуют алгоритмы, позволяющие проводить классификацию не только в пространстве признаков, измеренных с помощью метрических шкал, но и для шкал номинальных и порядковых.

Процедуры кластерного анализа распределяются по следующим направлениям.

1. *Иерархические классификации*, в результате которых получают схему взаимосвязи объектов или признаков в форме дендрограммы.

2. *Структурные классификации* предполагают предварительное определение центров сгущений объектов в пространстве. По мере присоединения к каждому центру конкретных наблюдений характеристик центров кластеров и их количество уточняется.

Регрессионный анализ[1]

Основная *цель регрессионного анализа* – возможность осуществления прогнозирования. Сначала для простоты изложения рассмотрим случай, когда у нас имеется только два признака – X и Y – и нас интересует зависимость между ними. Другими словами, сначала предположим, что наша "группа признаков" состоит из одного признака – X (потом перейдем к случаю, когда вместо одного X фигурируют несколько признаков). Мы знаем, что о связи_между признаками говорит соответствующий коэффициент корреляции: чем ближе значение модуля этого коэффициента к 1, тем более сильна эта связь, т.е. тем с большей уверенностью мы можем полагать, что с ростом значений одного признака растут (если коэффициент корреляции положителен) или убывают (если коэффициент корреляции отрицателен)

[1] Подробно см. Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. Киев: Наукова думка, 1982; Толстова Ю.Н. Анализ социологических данных. – М.: Научный мир, 2000. – С.2910-319.

значения другого (напомним, что коэффициент корреляции измеряет линейную связь между переменными; отметим, однако, что приводимые рассуждения справедливы и для других коэффициентов связи, например, для корреляционного отношения, дающего возможность оценить криволинейную связь). Но при этом мы совершенно не можем сказать о том, в какой степени возрастет значение *Y*, если значение *X* увеличится, скажем, на 1. А ситуации здесь могут быть весьма разными.

Итак, для того, чтобы делать прогноз о том, как изменится значение Y при том или ином изменении значения *X*, нам желательно знать, как говорят, форму связи между этими переменными, т.е. желательно найти функцию вида $Y = f(X)$. Подчеркнем, что отношение между *X* и *Y* несимметрично: речь идет именно о зависимости второй переменной от первой, именно о возможности прогноза значения *Y* от *X*, а не наоборот.

Поиск функции *f* предполагает разработку определенной модели связи между переменными, опирающуюся на априорные знания исследователя. Найденная с помощью регрессионной техники зависимость – это тоже некоторая модель реальности – модель, в соответствии с которой и находятся значения *Y* на основе информации о значениях признака *X*.

Вспомним, что в социологии мы имеем дело не с функциональными, а с корреляционными зависимостями, то есть одному значению *X* соответствует несколько значений *Y*. Тогда для изучения зависимости для каждого значения *X* рассчитывается среднее значение *Y* и изучается зависимость от *X* именно таких средних. Таким образом, необходимо найти функцию $\overline{Y}_X = f(X)$.

Фиксируя какое-либо значение *X*, равное, например, X_i (т.е. рассматривая совокупность объектов, обладающих этим значением), мы имеем дело с некоторым условным распределением *Y* (которое образуют значения зависимой переменной *Y*, вычисленные для объектов, обладающих значением X_i признака *X*). Ясно, что чем меньше разброс зависимого признака в условных распределениях, тем больше можно верить прогнозу значений этого признака, осуществляемому с помощью уравнения регрессии. Напротив, большой разброс может полностью лишить нас возможности делать прогноз: утверждение о том, что для такого-то X_i переменная *Y* в среднем равна соответствующему условному среднему, не будет иметь никакой практической ценности из-за того, что бессмысленным станет сам расчет средней величины.

Данный метод анализа был создан для анализа количественных данных. Использовать регрессионную технику для анализа номинальной шкалы бессмысленно. Для того чтобы на основе информации, полученной по номинальной шкале, можно было построить уравнение регрессии, эту информацию необходимо преобразовать. Соответствующее преобразование носит название дихотомизации номинальных данных. Этот подход применяется очень широко, поскольку его использование как бы “открывает дверь” для применения подавляющего большинства “количественных” методов с целью анализа номинальных данных. Для этого вместо каждого номинального признака, принимающего *к* значений, вводим *k* новых дихотомических (т.е. принимающих два значения, будем обозначать эти значения 0 и 1). Применение регрессионной техники к преобразованным номинальным данным называется номинальным регрессионным анализом.

Практическое задание

По выбранной Вами на первых занятиях проблеме укажите, какие виды многомерного анализа Вы сможете применить.

Сформулируйте цель применения данного метода, укажите переменные, которые будут подвергнуты данному виду анализа, предложите гипотезы о возможных результатах.

ВОПРОСЫ И ЗАДАНИЯ ДЛЯ КОНТРОЛЯ

Теоретические вопросы

1. Статистический подход в социологии.
2. Роль статистической закономерности в социологию.
3. Место этапа анализа данных в структуре социологического исследования.
4. Организация матрицы первичных данных.
5. Организация матрицы сгруппированных данных.
6. Виды анализа данных.
7. Основные понятия выборочного метода.
8. Виды выборочных исследований.
9. Расчет характеристик простой случайной выборки.
10. Одномерное распределение для номинальных шкал. Организация частотной таблицы. Расчет различных видов процентов. Расчет показателей центра распределения и вариации. Графическое изображение.
11. Одномерное распределение для порядковых шкал. Особенности построения таблицы. Расчет показателей центра распределения и вариации. Использование условных средних (индексов).
12. Одномерное распределение для метрических шкал. Организация таблицы распределения. Расчет показателей центра распределения и вариации.
13. Понятие взаимосвязи, виды взаимосвязи.
14. Логика проверки статистических гипотез о взаимосвязи двух переменных.
15. Случаи двухмерного распределения, когда зависимая переменная является номинальной. Построение таблиц распределения. Проверка статистической значимости взаимосвязи. Оценка силы взаимосвязи.
16. Случаи двухмерного распределения, когда зависимая переменная является порядковой (без расчета условного индекса). Общее и особенное для случаев двухмерного распределения с зависимой порядковой переменной. Коэффициенты ранговой корреляции. Коэффициент Спирмена. Коэффициент Кендалла. γ-коэффициент.
17. Анализ взаимосвязи, когда зависимая переменная является количественной.
18. Многомерный анализ и природа социальных взаимосвязей.
19. Детерминационный анализ: основные понятия, этапы реализации процедуры, интерпретация результатов, ограничения.

20. Регрессионный анализ: основные понятия, этапы реализации процедуры, интерпретация результатов, ограничения.

Самостоятельное практическое задание по курсу

Задание выполняется с целью приобрести навыки принятия решения о необходимых для изучения конкретной исследовательской проблемы методах статистического анализа и представляет собой письменную работу, содержащую следующие элементы:

1. Выбор исследовательской проблемы. Тема исследования.
2. Исследовательские гипотезы.
3. Система индикаторов для проверки каждой гипотезы.
4. Переменные и шкалы, по которым эти переменные могут быть замерены.
5. Определение типа каждой переменной.
6. Статистические показатели, которые необходимо рассчитать по каждой переменной и цель их расчета.
7. Статистические гипотезы относительно связи между двумя переменными.
8. Алгоритм проверки этих гипотез.
9. Необходимые вторичные переменные, сформулированные различными методами.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Основной

1. Девятко, И.Ф. Методы социологического исследования / И.Ф. Девятко. – М.: Университет, 2002. – 208 с.
2. Татарова, Г.Г. Методология анализа данных в социологии / Г.Г. Татарова. – М.: Изд. дом «Стратегия», 1998. – 222 с.
3. Толстова, Ю.Н. Анализ социологических данных / Ю.Н. Толстова. – М.: Научный мир, 2000. – 352 с.
4. Толстова, Ю. Н. Социология и математика / Ю.Н. Толстова. – М.: Научный мир, 2003. – 323 с.
5. Хили, Д. Статистика: социологические и маркетинговые исследования / Д. Хили. – Киев и др.; СПб. и др.: Питер: DiaSoft, 2005. – 637 с.

Дополнительный

6. Девятко, И. Ф. Диагностическая процедура в социологии: Очерк истории и теории / И.Ф. Девятко. – М.: Наука, 1993–168 с.
7. Елисеева, И. И. Группировка, корреляция, распознавание образов: Стат. методы классификации и измерения связей / И.И. Елисеева. – М.: Статистика, 1977. – 143 с.
8. Елисеева, И. И. Логика прикладного статистического анализа / И.И. Елисеева. – М.: Финансы и статистика, 1982. – 192 с.
9. Елисеева, И. И. Статистические методы измерения связей / И.И. Елисеева. – Л.: Изд-во ЛГУ, 1982. – 134 с.
10. Интерпретация и анализ данных в социологических исследованиях / В.Г. Андреенков, Ю.Н. Толстова, И.И. Елисеева и др.. – М.: Наука, 1987. – 254 с.
11. Клигер, С. А. и др. Шкалирование при сборе и анализе социологической информации / С.А. Клигер, М.С. Косолапов, Ю.Н. Толстова. – М.: Наука, 1978. – 112 с.
12. Математические методы анализа и интерпретация социологических данных / В. Г. Андреенков, К.Д. Аргунова, В.И. Паниотто и др.. – М.: Наука, 1989. – 175 с.

13. Паниотто, В. И. Качество социологической информации: Методы оценки и процедуры обеспечения / В.И. Паниотто. – Киев: Наукова думка, 1986. – 206 с.
14. Паниотто, В.И. Количественные методы в социологических исследованиях / В.И. Паниотто, В.С. Максименко. – Киев: Наукова думка, 1982. – 272 с.
15. Рабочая книга социолога / Под ред. Г.В. Осипова. – М.: Наука, 1983. – 477 с.
16. Типология и классификация в социологических исследованиях / В.Г. Андреенков, Ю.Н. Толстова, А.А. Мирзоев и др.. – М.: Наука, 1982. – 296 с.
17. Толстова, Ю. Н. Логика математического анализа социологических данных / Ю.Н. Толстова. – М.: Наука, 1991. – 110 с.
18. Экспертные оценки в социологических исследованиях / С.Б. Крымский, Б. Б. Жилин, В. И. Паниотто и др. – Киев: Наукова думка, 1990. – 318 с.
19. Ядов, В.А. Стратегия социологического исследования: Описание, объяснение, понимание социальной реальности / В.А. Ядов. – М.: Академкнига: Добросвет, 2003. – 595 с.

ПРИЛОЖЕНИЕ

Таблица 1

Критические значения для χ^2 распределения[1]

Различия между двумя распределениями могут считаться достоверными, если $\chi^2_{эмп}$ достигает или превышает $\chi^2_{0,05}$.

df	Уровень значимости *p*		df	Уровень значимости *p*		df	Уровень значимости *p*	
	0,05	0,01		0,05	0,01		0,05	0.01
1	3,841	6,635	31	44,985	52,191	61	80,232	89,591
2	5,991	9,210	32	46,194	53,486	62	81,381	90,802
3	7,815	11,345	33	47,400	54,776	63	82,529	92,010
4	9,488	13,277	34	48,602	56,061	64	83,675	93,217
5	11,070	15,086	35	49,802	57,342	65	84,821	94,422
6	12,592	16,812	36	50,998	58,619	66	85,965	95,626
7	14,067	18,475	37	52,192	59,892	67	87,108	96,828
8	15,507	20,090	38	53,384	61,162	68	88,250	98,028
9	16,919	21,666	39	54,572	62,428	69	89,391	99,227
10	18,307	23,209	40	55,758	63,691	70	90,631	100,425
11	19,675	24,725	41	56,942	64,950	71	91,670	101,621
12	21,026	26,217	42	58,124	66,206	72	92,808	102,816
13	22,362	27,688	43	59,304	67,459	73	92,945	104,010
14	23,685	29,141	44	60,481	68,709	74	95,081	105,202
15	24,996	30,578	45	61,656	69,957	75	96,217	106,393
16	26,296	32,000	46	62,830	71,201	76	97,351	107,582
17	27,587	33,409	47	64,001	72,443	77	98,484	108,771
18	28,869	34,805	48	65,171	73,683	78	99,617	109,958
19	30,144	36,191	49	66,339	74,919	79	100,749	111,144
20	31,410	37,566	50	67,505	76,154	80	101,879	112,329
21	32,671	38,932	51	68,669	77,386	81	103,010	113,512
22	33,924	40,289	52	69,832	78,616	82	104,139	114,695
23	35,172	41,638	53	70,993	79,843	83	105,267	115,876
24	36,415	42,980	54	72,153	81,069	84	106,395	117,057
25	37,652	44,314	55	73,311	82,992	85	107,522	118,236
26	38,885	45,642	56	74,468	83,513	86	108,648	119,414
27	40,113	46,963	57	75,624	84,733	87	109,773	120,591
28	41,337	48,278	58	76,778	85,950	88	110,898	121,767
29	42,557	49,588	59	77,931	87,166	89	112,022	122,942
30	43,773	50,892	60	79,082	88,379	90	113,145	124,116

[1] По: Сидоренко, Е.В. Методы математической обработки в психологии. – СПб.: Социально-психологический центр, 1996. – 350 с.

Таблица 2

Таблица значений критических точек стандартного нормального распределения для различных уровней значимости[1]

Вероятность ошибки α	0,01	0,025	0,05	0,10	0,20	0,30
Z критическое	2,3263	1,9600	1,6449	1,2816	0,8416	0,5255

Таблица 3

Критические значения коэффициента ранговой корреляции Спирмена[2]

Количество	Вероятность ошибки			Количество	Вероятность ошибки		
пар рангов	0,05	0,01	0,001	пар рангов	0,05	0,01	0,001
6	0,829	1,000	—	25	0,398	0,510	0,618
7	0,745	0,893	1,000	30	0,362	0,466	0 570
8	0,691	0,857	0,952	35	0,333	0,429	0,534
9	0,683	0,817	0,917	40	0,311	0,402	0,501
10	0,636	0,782	0,891	45	0,294	0,380	0,475
11	0,618	0,754	0,867	50	0,279	0,361	0,450
12	0,580	0,727	0,823	60	0,254	0,330	0,415
13	0,555	0,698	0,801	70	0,235	0,306	0,385
14	0,534	0,675	0,793	80	0,220	0,286	0,361
15	0,518	0,654	0,760	90	0,207	0,270	0,341
16	0,500	0,632	0,741	100	0,196	0,257	0,324
17	0,485	0,615	0,724	150	0,160	0,209	0,265
18	0,472	0,598	0,709	200	0,139	0,182	0,231
19	0,458	0,583	0,694	500	0,087	0,115	0,148
20	0,445	0,568	0,679	1000	0,062	0,081	0,104

Таблица 4

Таблица значений функции Лапласа при разных значениях t[3]

Ф	0,68269	0,95000	0,95450	0,99730
t	1	1,96	2	3

Таблица 5

Критические значения для *t*-распределения Стьюдента[4]

Объем совокупности	Вероятность ошибки				
	0,20	0,10	0,05	0,02	0,01
30	1,310	1,697	2,042	2,457	2,750
∞	1,282	1,645	1,960	2,326	2,576

[1] По: Рабочая книга социолога. – М.: Едиториал УРСС, 2003. – С. 451.

[2] По: Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. – Киев: Наукова думка, 1982. – С. 259.

[3] По: Ефимова, М.Р., Петрова Е.В., Румянцева, В.Н. Общая теория статистики. – М.: ИНФРА-М, 1998. – С. 394-396.

[4] По: Рабочая книга социолога. – М.: Едиториал УРСС, 2003. – С. 452.

Вероника Юлиевна Колчинская

Анализ данных в социологии

Учебное пособие

Техн. редактор А.В. Миних

Издательство Южно-Уральского государственного университета

Подписано в печать 28.12.2006. Формат 60х84 1/16. Усл. печ. л. 5,11.

Уч.-изд. л. 4,86. Тираж 100 экз. Заказ 522. Цена С.

Отпечатано в типографии Издательства ЮУрГУ. 454080, г. Челябинск, пр. им. В.И. Ленина, 76